

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

REMARKS

Applicants request reconsideration of the rejections, and allowance of the claims of the above-captioned patent Application in view of the following remarks. Claims 1-3, 6-18, 20, 23, and 28-29 are currently pending as the claims for which Applicants have paid, of which claims 1, 2, 17, and 18 are presently being prosecuted. Claims 4, 5, 19, 21-22, 24-27, and 30-45 have been canceled.

SUMMARY OF THE INVENTION

Applicants' invention is directed to a human integral membrane protein ("IMP"), which has strong similarity to human stomatin, and is a member of the vertebrate stomatin gene superfamily. Human stomatin is a membrane protein which is widely expressed in erythrocytes and ciliated cells. IMP is 356 amino acids in length. As shown in Figures 2A, 2B, and 2C IMP has chemical and structural homology with human stomatin (GI 31069; SEQ ID NO:3) and *C. elegans* MEC-2 protein (GI 10654523; SEQ ID NO:4). In particular, residues 79-209 of IMP are similar to residues 96-226 of stomatin (33% identity, 60% similarity) and residues 218-253 of IMP are strongly similar to residues 211-246 of stomatin (30% identity, 52% similarity). In addition, IMP contains numerous potential phosphorylation sites (*i.e.*, typically the hydroxyl groups of serine, threonine and tyrosine residues although asparagine, histidine and lysine residues may also be phosphorylated), including potential sites for phosphorylation by cAMP-dependent protein kinase (*e.g.*, R-X-S/T) (*i.e.*, S₂₉, T₃₆, S₁₅₂ and S₂₁₃).

Northern analysis (Figure 3) shows the expression of SEQ ID NO:1 in various libraries, at least 38% of which are cancerous or immortalized. Of particular note is the expression of IMP mRNA in prostate tumor (2/13), breast tumor (1/13), and pancreatic tumor libraries (1/13). This pattern of expression demonstrates that IMP serves as a marker for cancerous cells, particularly prostate tumor cells.

As such, the polynucleotide of the claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which necessarily require detailed knowledge of how the polypeptide coded for by the polynucleotide works. As a result of the benefits of these uses, the claimed invention already enjoys significant commercial success.

Claims 1-2 and 17-18 stand rejected under 35 U.S.C. § 101 based on the allegation that the claimed invention lacks patentable utility. The rejection alleges in particular that:

- there is no well-established, specific and substantial utility for the claimed polypeptide and fragments/variants thereof as there is no disclosure of any correlation between a specific disorder and an altered level of expression of the claimed polypeptide, and thus the results of a gene expression monitoring assay would have no meaning (Final Office Action mailed August 1, 2003; pp. 5-6).
- the claimed invention is not supported by either a substantial and specific asserted utility or a well established utility. The specification discloses no uses for the broadly claimed polypeptides. A specific utility is one that is particular to the subject matter claimed, while a substantial utility is one that defines a “real world” use. Utilities that require or constitute carrying out further research to identify or reasonably confirm a “real world” context of use are not substantial utilities (Final Office Action mailed August 1, 2003 at pp. 6-7).

Utility Rejection – 35 USC § 101

The rejection of claims 1, 2, 17 and 18 is improper, as the inventions of those claims have a patentable utility as set forth in the specification, and/or a utility well-known to one of ordinary skill in the art.

The invention at issue, identified in the patent application as novel integral membrane protein, abbreviated as IMP, is a polypeptide sequence encoded by a gene that is expressed in prostate tumor, breast tumor, and pancreatic tumor tissues of humans. The novel polypeptide is demonstrated in the specification to be a member of the class of integral membrane proteins, whose biological functions include the regulation of ion channel activity (Specification at p. 3, lines 12-21). As such, the claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which require knowledge of how the polypeptide actually functions. The claimed invention also can be used as tissue or tumor marker (Specification at p. 15, lines 12-16).

The similarity of the claimed polypeptide to another polypeptide of known, undisputed utility by itself demonstrates utility beyond the reasonable probability required by law. IMP is, in that regard,

homologous to stomatin-like proteins. Though not necessary to demonstrate the utility of the claimed SEQ ID NO:1 polypeptide, Applicants respectfully direct the Examiner's attention to the enclosed paper by Wang et al., Identification and characterization of human SLP-2, a novel homologue of stomatin (Band 7.2b) present in erythrocytes and other tissues, Journal of Biological Chemistry 2000;275(11):8062-8071 (Enclosure No. 1). This post-filing reference describes the characterization of stomatin-like proteins SLP-1 and SLP-2; of which, SLP-2 has been implicated in the control of ion channel permeability, mechanoreception, and lipid domain organization. Further, SLP-2 is 95% identical over 358 amino acid residues to SEQ ID NO:1 (Enclosure No. 2; alignment with AF190167). This post-filing reference corroborates Applicants' prior identification of SEQ ID NO:1 as a stomatin protein. In addition, Applicants respectfully direct the Examiner's attention to the enclosed paper by Owczarek C.M. et al., A novel member of the STOMATIN/EPB72/mec-2 family, stomatin-like 2 (STOML2), is ubiquitously expressed and localized to HSA chromosome 9p13.1, Cytogenet Cell Genet. 2001;92(3-4):196-203 (Enclosure No. 3). This post-filing reference describes stomatin-like 2 (STOML2), which has been mapped to a chromosomal region that is rearranged in some cancers and thought to contain a gene responsible for acromesomelic dysplasia. In addition, STOML2 is 96% identical over 356 amino acid residues to SEQ ID NO:1 (Enclosure No. 4; alignment with AF282596). Thus, this post-filing reference also corroborates Applicants' prior identification of SEQ ID NO:1 as a stomatin protein. Applicants also respectfully direct the Examiner's attention to the enclosed paper by Fricke et al., Stomatin immunoreactivity in ciliated cells of the human airway epithelium, Anat Embryol (Berl), 2003 Jul;207(1):1-7, Epub, May 21, 2003 (Enclosure No. 5), wherein it is stated that stomatin is situated with cholesterol+sphingomyelin-rich "rafts" in the cytomembrane, such that it could be a candidate for a membrane-associated mechanotransducer with a role in the control of ciliary motility. Further, the authors indicate that stomatin, as a raft protein, might be a microtubule associated protein moving along the outer surface of the microtubules to its terminal site of action in the cilia. The presence of stomatin in microvilli supports the hypothesis of a co-localization with beta- and gamma- ENaC and, in conclusion, their potential functional interaction to control the composition of periciliary mucus electrolytes. While stomatin is absent in erythrocytes from patients with hereditary stomatocytosis, linkage studies have established that stomatin is not the cause of this disorder. However,

characterization of members of the stomatin gene family may lead to further understanding of certain hemolytic conditions. Accordingly, these alignments demonstrate the successful application of homology-based characterization of newly identified proteins, such as IMP.

As shown in the Specification, IMP has chemical and structural homology with human stomatin (GI 31069; SEQ ID NO:3) and *C. elegans* MEC-2 protein (GI 10654523; SEQ ID NO:4). In particular, residues 79-209 of IMP are similar to residues 96-226 of stomatin (33% identity, 60% similarity) and residues 218-253 of IMP are strongly similar to residues 211-246 of stomatin (30% identity, 52% similarity). This is more than enough homology to demonstrate a reasonable probability that the utility of SLP2 and STOML2 can be imputed to the claimed invention. It is well-known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. Brenner et al., Proc. Natl. Acad. Sci. U.S.A. 95:6073-78 (1998) (Enclosure No. 6). Given homology in excess of 40% over many more than 70 amino acid residues, the probability that the claimed polypeptide is related to SLP-2 and STOML2 is, accordingly, very high.

Furthermore, the fact that the claimed polypeptide is a member of the integral membrane protein family alone demonstrates utility. Each of the members of this class, regardless of their particular functions, are useful. There is no evidence that any member of this class of polypeptides, let alone a substantial number of them, would not have some patentable utility. It follows that there is a more than substantial likelihood that the claimed polypeptide also has patentable utility, regardless of its actual function. The law has never required a patentee to prove more.

There is, in addition, direct proof of the utility of the claimed invention. Applicants have submitted the Declaration of Furness describing some of the practical uses of the claimed invention in gene and protein expression monitoring applications as they would have been understood at the time of the patent application. The Furness Declaration describes, in particular, how the claimed polypeptide can be used in protein expression analysis techniques such as 2-D PAGE gels and western blots. Using the claimed invention with these techniques, persons of ordinary skill in the art can better assess, for example, the potential toxic effect of a drug candidate. (Furness Declaration ¶12 (b)).

The Examiner contends that the claimed polypeptide cannot be useful without precise knowledge of its function. But the law has never required knowledge of biological function to prove utility. It is the claimed invention's uses, not its functions, that are the subject of a proper analysis under the utility requirement.

As demonstrated by the Furness Declaration, the person of ordinary skill in the art can achieve beneficial results from the claimed polypeptide in the absence of any knowledge as to the precise function of the protein. The uses of the claimed polypeptide for gene expression monitoring applications including toxicology testing are in fact independent of its precise function.

The Office Action is replete with arguments made and positions taken in a misplaced attempt to justify the rejections of the claims under 35 U.S.C. §§ 101 and 112. The Examiner's positions and arguments include that the invention (i.e., the SEQ ID NO:1 polypeptide) is not supported by "credible, specific and substantial utility of human synthase [*sic:stomatin*] activity" (Final Office Action of August 1, 2003; p. 5).

The Declaration of Furness under 37 C.F.R. § 1.132 (the Furness Declaration) shows the many substantial reasons why the Examiner's positions and arguments with respect to the use of the SEQ ID NO:1 polypeptide are without merit.

I. The Applicable Legal Standard

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

An invention is "useful" under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) ("to violate Section 101 the claimed device must be totally incapable of achieving a useful result"); *Fuller v. Berger*, 120 F. 274, 275 (7th

Cir. 1903) (test for utility is whether invention “is incapable of serving any beneficial end”).

Juicy Whip Inc. v. Orange Bang Inc., 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: “[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility.” *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. See *Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a “nebulous expression” such as “biological activity” or “biological properties” that does not convey meaningful information about the utility of what is being claimed. *Cross v. Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be “substantial.” *Brenner*, 383 U.S. at 534. A “substantial” utility is a practical, “real-world” utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a “well-established” utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examination Procedure at § 706.03(a). Only if there is no “well-established” utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*, 51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the

burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the Patent Office must provide evidence or sound scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a “substantial likelihood” of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

II. Toxicology testing, use as a tissue or tumor marker, regulation of membrane conductance, and regulation of ion channel activity are sufficient utilities under 35 U.S.C. §§ 101 and 112, first paragraph

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are “well-established” uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application’s specification. These uses are explained, in detail, in the Furness Declaration accompanying this response. Objective evidence, not considered by the Patent Office, further corroborates the credibility of the asserted utilities.

A. The claimed polypeptide’s membership in the integral membrane protein family demonstrates utility

Because there is a substantial likelihood that the claimed IMP is a member of the family of polypeptides known as integral membrane proteins, the members of which are indisputably useful, there is by implication a substantial likelihood that the claimed polypeptide is similarly useful. Applicants need not show any more to demonstrate utility. *In re Brana*, 51 F.3d at 1567.

It is undisputed that the claimed polypeptide is a protein having the sequence shown as SEQ ID NO:1 in the patent application and referred to as IMP in that application. Applicants have demonstrated by more than reasonable probability that IMP is a member of the integral membrane protein family, and that the integral membrane protein of proteins includes stomatin and stomatin-like

proteins, each of which regulate membrane conductance. IMP has structural and chemical homology with stomatin.

The Examiner must accept Applicants' assertion that the claimed polypeptide is a member of integral membrane protein family and that utility is credible to a reasonable probability unless the Examiner can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt utility. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not provided sufficient evidence or sound scientific reasoning to the contrary.

Nor has the Examiner provided any evidence that any member of the integral membrane protein family, let alone a substantial number of those members, is not useful. In such circumstances the only reasonable inference is that the claimed polypeptide must be, like the other members of the integral membrane protein family, useful.

B. The Office Action failed to demonstrate that a person of ordinary skill in the art would reasonably doubt the utility of the claimed invention

Based principally on citations to scientific literature identifying some of the difficulties involved in predicting protein function, the Office Action rejected the pending claims on the ground that the Applicants cannot impute utility to the claimed invention based on the homology between the encoded polypeptide, IMP, and another polypeptide. The Office Action's rejection is both incorrect as a matter of fact and as a matter of procedural law.

While the Office Action has cited literature identifying some of the difficulties that may be involved in predicting protein function, none suggests that functional homology cannot be inferred by a reasonable probability in this case. Importantly, none contradicts Bork's later findings that there is a 70% accuracy rate for bioinformatics-based predictions in general, and a 90% accuracy rate for the prediction of functional features by homology. Bork, *Genome Research* 10:398-400 (2000). At most, these articles individually and together stand for the proposition that it is difficult to make predictions about function with certainty. The standard applicable in this case is not, however, proof to certainty, but rather proof to reasonable probability.

The literature cited in the Office Action may show that Applicants cannot prove function by homology with **certainty**, but Applicants need not meet such a rigorous standard of proof. Under the applicable law, once the Applicants demonstrate a *prima facie* case of homology, the Office must accept the assertion of utility to be true unless the Office comes forward with evidence showing a person of ordinary skill would doubt the asserted utility could be achieved by a reasonable probability. See *In re Brana*, 51 F.3d at 1566; *In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Office has not made such a showing and, as such, the Office Action's rejection should be withdrawn.

C. The uses of IMP for toxicology testing, drug discovery, and disease diagnosis are practical uses that confer “specific benefits” to the public

The claimed invention has specific, substantial, real-world utility by virtue of its use in toxicology testing, drug development and disease diagnosis through gene expression profiling. These uses are explained in detail in the Furness Declaration, the substance of which is not rebutted by the Patent Examiner. There is no dispute that the claimed invention is in fact a useful tool in two-dimensional polyacrylamide gel electrophoresis (“2-D PAGE”) analysis and western blots used to monitor protein expression and assess drug toxicity.

The instant application is a continuation application of and claimed priority to United States patent application Serial No. 09/095,351 filed on June 9, 1998 (hereinafter “the Hillman ‘351 application”), which in turn was a divisional application of and claimed priority to United States patent application Serial No. 08/781,562 filed on January 9, 1997 (hereinafter “the Hillman ‘562 application”), all having the identical specification.

In his Declaration, Mr. Furness explains the many reasons why a person skilled in the art who read the Hillman ‘562 application on January 9, 1997 would have understood that application to disclose the claimed polypeptide to be useful for a number of gene and protein expression monitoring applications, *e.g.*, in 2-D PAGE technologies, in connection with the development of drugs and the monitoring of the activity of such drugs. (Furness Declaration at, *e.g.*, ¶¶11-13). Much, but not all, of

Mr. Furness' explanation concerns the use of the claimed polypeptide in the creation of protein expression maps using 2-D PAGE.

2-D PAGE technologies were developed during the 1980's. Since the early 1990's, 2-D PAGE has been used to create maps showing the differential expression of proteins in different cell types or in similar cell types in response to drugs and potential toxic agents. Each expression pattern reveals the state of a tissue or cell type in its given environment, *e.g.*, in the presence or absence of a drug. By comparing a map of cells treated with a potential drug candidate to a map of cells not treated with the candidate, for example, the potential toxicity of a drug can be assessed. (Furness Declaration at ¶11.)

The claimed invention makes 2-D PAGE analysis a more powerful tool for toxicology and drug efficacy testing. A person of ordinary skill in the art can derive more information about the state or states or tissue or cell samples from 2-D PAGE analysis with the claimed invention than without it. As Mr. Furness explains:

In view of the Hillman '562 application, the Wilkins article, and other related pre-January 1997 publications, persons skilled in the art on January 9, 1997 clearly would have understood the Hillman '562 application to disclose the SEQ ID NO:1 polypeptide or the antibody to SEQ ID NO:1 polypeptide to be useful in 2-D PAGE analyses for the development of new drugs and monitoring the activities of drugs for such purposes as evaluating their efficacy and toxicity (Furness Declaration, ¶10)

* * *

Persons skilled in the art would appreciate that a 2-D PAGE map that utilized the SEQ ID NO:1 polypeptide sequence would be a more useful tool than a 2-D PAGE map that did not utilize this protein sequence in connection with conducting protein expression monitoring studies on proposed (or actual) drugs for treating cancers for such purposes as evaluating their efficacy and toxicity. (Furness Declaration, ¶12)

Mr. Furness' observations are confirmed in the literature published before the filing of the patent application. Wilkins, for example, describes how 2-D gels are used to define proteins present in various tissues and measure their levels of expression, the data from which is in turn used in databases:

For proteome projects, the aim of [computer-aided 2-D PAGE] analysis . . . is to catalogue all spots from the 2-D gel in a qualitative and if possible quantitative manner, so as to define the number of proteins present and their levels of expression. Enclosure

gel images, constructed from one or more gels, for the basis of two-dimensional gel databases. (Wilkins, Tab 1, p. 26).

D. The use of proteins expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now “well-established”

The technologies made possible by expression profiling using polypeptides are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, as described by Furness in his Declaration.

Toxicology testing is now standard practice in the pharmaceutical industry. See, e.g., John C. Rockett, et. al., Differential gene expression in drug metabolism and toxicology: practicalities, problems, and potential, *Xenobiotica* 29:655-691 (July 1999) (Enclosure No. 7):

Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. ((Enclosure No. 7), p. 656)

To the same effect are several other scientific publications, including Emile F. Nuwaysir, et al., Microarrays and Toxicology: The Advent of Toxicogenomics, *Molecular Carcinogenesis* 24:153-159 (1999) (Enclosure No. 8); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, *Toxicology Letters* 112-13:467-471 (2000) (Enclosure No. 9).

The more genes – and, accordingly, the polypeptides they encode -- that are available for use in toxicology testing, the more powerful the technique. Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator of the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding (Enclosure No. 10) Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

In fact, the potential benefit to the public, in terms of lives saved and reduced health care costs, are enormous. Recent developments provide evidence that the benefits of this information are already beginning to manifest themselves. Examples include the following:

- In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter gene, and chromosomal mapping location, to identify the key gene associated with Tangier disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.
- In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte's genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.
- In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

Because the Patent Examiner failed to address or consider the "well-established" utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner's rejections should be withdrawn regardless of their merit.

E. Objective evidence corroborates the utilities of the claimed invention

There is in fact no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a "real-world" utility exists. "Real-world" evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or

entity other than the patentee is conclusive proof of utility. *United States Steel Corp. v. Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a vibrant market has developed for databases containing all expressed genes (along with the polypeptide translations of those genes). (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable). The databases sold by Applicants' assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations. Incyte sells its database containing the claimed sequence and millions of other sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte's customers and the scientific community have acknowledged that Incyte's databases have proven to be valuable in, for example, the identification and development of drug candidates. As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte's discovery of the claimed polypeptide, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

III. The Patent Examiner's Rejections Are Without Merit

Rather than responding to the evidence demonstrating utility, the Examiner attempts to dismiss it altogether by arguing that the disclosed and well-established utilities for the claimed polypeptide are not "specific" or "well-established" utilities (Final Office Action at p. 3). The Examiner is incorrect both as a matter of law and as a matter of fact.

A. The Precise Biological Role Or Function Of An Expressed Polypeptide Is Not Required To Demonstrate Utility

The Patent Examiner's primary rejection of the claimed invention is based on the ground that, without information as to the precise biological role ("what IMP is, how it functions," as well as its "relationship to any specific disease" (Office Action at p. 3)) of the claimed invention, the claimed invention's utility is not sufficiently specific.

It may be that specific and substantial interpretations and detailed information on biological function are necessary to satisfy the requirements for publication in some technical journals, but they are not necessary to satisfy the requirements for obtaining a United States patent. The relevant question is not, as the Examiner would have it, whether it is known how or why the invention works, *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999), but rather whether the invention provides an “identifiable benefit” in presently available form. *Juicy Whip Inc. v. Orange Bang Inc.*, 185 F.3d 1364, 1366 (Fed. Cir. 1999). If the benefit exists, and there is a substantial likelihood the invention provides the benefit, it is useful. There can be no doubt, particularly in view of the Furness Declaration (at, e.g., ¶¶ 10-13), that the present invention meets this test.

The threshold for determining whether an invention produces an identifiable benefit is low. *Juicy Whip*, 185 F.3d at 1366. Only those utilities that are so nebulous that a person of ordinary skill in the art would not know how to achieve an identifiable benefit and, at least according to the PTO guidelines, so-called “throwaway” utilities that are not directed to a person of ordinary skill in the art at all, do not meet the statutory requirement of utility. Utility Examination Guidelines, 66 Fed. Reg. 1092 (Jan. 5, 2001).

Knowledge of the biological function or role of a biological molecule has never been required to show real-world benefit. In its most recent explanation of its own utility guidelines, the PTO acknowledged as much (66 F.R. at 1095):

[T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have specific and substantial utility because, e.g., it hybridizes near a disease-associated gene or it has gene-regulating activity.

By implicitly requiring knowledge of biological function for any claimed polypeptide, the Examiner has, contrary to law, elevated what is at most an evidentiary factor into an absolute requirement of utility. Rather than looking to the biological role or function of the claimed invention, the Examiner should have looked first to the benefits it is alleged to provide.

B. Membership in a Class of Useful Products Can Be Proof of Utility

Despite the uncontradicted evidence that the claimed polypeptide is a member of the integral membrane protein family, whose members indisputably are useful, the Examiner refused to impute the utility of the members of the integral membrane protein family to IMP. In the Office Action at p.6, the Patent Examiner takes the position that unless Applicants can identify which particular biological function within the class of integral membrane protein is possessed by IMP, utility cannot be imputed. To demonstrate utility by membership in the class of integral membrane proteins, the Examiner would require that all integral membrane proteins possess a “common” utility.

There is no such requirement in the law. In order to demonstrate utility by membership in a class, the law requires only that the class not contain a substantial number of useless members. So long as the class does not contain a substantial number of useless members, there is sufficient likelihood that the claimed invention will have utility and a rejection under 35 U.S.C. § 101 is improper. That is true regardless of how the claimed invention ultimately is used and whether the members of the class possess one utility or many. *See Brenner v. Manson*, 383 U.S. 519, 532 (1966); *Application of Kirk*, 376 F.2d 936, 943 (CCPA 1967).

Membership in a “general” class is insufficient to demonstrate utility only if the class contains a substantial number of useless members. There would be, in that case, a substantial likelihood that the claimed invention is one of the useless members of the class. In the few cases in which class membership did not prove utility by substantial likelihood, the classes did in fact include predominately useless members. *E.g.*, *Brenner* (man-made steroids); *Kirk* (same); *Natta* (man-made polyethylene polymers).¹

The Examiner addresses IMP as if the general class in which it is included is not the integral membrane protein family, but rather all polypeptides, including the vast majority of useless theoretical molecules not occurring in nature, and thus not pre-selected by nature to be useful. While these

¹At a recent Biotechnology Customer Partnership Meeting, PTO Senior Examiner James Martinell described an analytical framework roughly consistent with this analysis. He stated that when an Appellants’ claimed protein “is a member of a family of proteins that already are known based upon sequence homology,” that can be an effective assertion of utility.

“general classes” may contain a substantial number of useless members, the integral membrane protein family does not. The integral membrane protein family is sufficiently specific to rule out any reasonable possibility that IMP would not also be useful like the other members of the family.

Because the Examiner has not presented any evidence that the integral membrane protein class of proteins has any, let alone a substantial number, of useless members, the Examiner must conclude that there is a “substantial likelihood” that the IMP encoded by the claimed polypeptide is useful.

Even if the Examiner’s “common utility” criterion were correct – and it is not – the integral membrane protein family would meet it. It is undisputed that known members of the integral membrane protein family regulate membrane conductance and regulate ion channel activity. A person of ordinary skill in the art need not know any more about how the claimed invention regulates membrane conductance to use it, and the Examiner presents no evidence to the contrary. Instead, the Examiner makes the conclusory observation that a person of ordinary skill in the art would need to know whether, for example, any given integral membrane protein regulates membrane conductance. The Examiner then goes on to assume that the only use for IMP absent knowledge as to how this member of the integral membrane protein family actually works is further study of IMP itself.

Not so. As demonstrated by Applicants, knowledge that IMP is an integral membrane protein is more than sufficient to make it useful for the diagnosis and treatment of various cancers. Indeed, IMP has been shown to be expressed in prostate, breast, and pancreatic tumor tissue libraries. The Examiner must accept these facts to be true unless the Examiner can provide evidence or sound scientific reasoning to the contrary. But the Examiner has not done so.

C. The uses of IMP in toxicology testing, drug discovery, and disease diagnosis are practical uses beyond mere study of the invention itself

To the extent that the Examiner rejected the claims at issue on the ground that the use of an invention as a tool for research is not a “substantial” use, the Examiner’s rejection assumes a substantial overstatement of the law, and is incorrect in fact. Therefore, it must be withdrawn.

There is no authority for the proposition that use as a tool for research is not a substantial utility. Indeed, the Patent Office itself has recognized that just because an invention is used in a research setting

does not mean that it lacks utility (Section 2107.01 of the Manual of Patent Examining Procedure, 8th Edition, August 2001, under the heading I. Specific and Substantial Requirements, Research Tools):

Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact “useful” in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified substantial utility and inventions whose asserted utility requires further research to identify or reasonably confirm.

The PTO’s actual practice has been, at least until the present, consistent with that approach. It has routinely issued patents for inventions whose only use is to facilitate research, such as DNA ligases, acknowledged by the PTO’s Training Materials to be useful.

The subset of research uses that are not “substantial” utilities is limited. It consists only of those uses in which the claimed invention is to be an **object** of further study, thus merely inviting further research on the invention itself. This follows from *Brenner*, in which the U.S. Supreme Court held that a process for making a compound does not confer a substantial benefit where the only known use of the compound was to be the object of further research to determine its use. *Id.* at 535. Similarly, in *Kirk*, the Court held that a compound would not confer substantial benefit on the public merely because it might be used to synthesize some other, unknown compound that would confer substantial benefit. *Kirk*, 376 F.2d at 940, 945. (“What Applicants are really saying to those in the art is take these steroids, experiment, and find what use they do have as medicines.”) Nowhere do those cases state or imply, however, that a material cannot be patentable if it has some other, additional beneficial use in research.

Such beneficial uses beyond studying the claimed invention itself have been demonstrated, in particular those described in the Furness Declaration. The Furness Declaration demonstrates that the claimed invention is a tool, rather than an object, of research, and it demonstrates exactly how that tool is used. Without the claimed invention, it would be more difficult to generate information regarding the properties of tissues, cells, drug candidates and toxins apart from additional information about the polypeptide itself.

D. The Patent Examiner Failed to Demonstrate That a Person of Ordinary Skill in the Art Would Reasonably Doubt the Utility of the Claimed Invention

Based principally on citations to scientific literature identifying some of the difficulties involved in predicting protein function, the Examiner rejected the pending claims on the ground that the applicant cannot impute utility to the claimed invention based on its structural similarity to another polypeptide undisputed by the Examiner to be useful. The Examiner's rejection is both incorrect as a matter of fact and as a matter of procedural law.

As demonstrated in § II.A., *supra*, the literature cited by the Examiner is not inconsistent with the Applicants' proof of homology by a reasonable probability. It may show that Applicants cannot prove function by homology with **certainty**, but Applicants need not meet such a rigorous standard of proof. Under the applicable law, once the applicant demonstrates a *prima facie* case of homology, the Examiner must accept the assertion of utility to be true unless the Examiner comes forward with evidence showing a person of ordinary skill would doubt the asserted utility could be achieved by a reasonable probability. See *In re Brana*, 51 F.3d at 1566; *In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not made such a showing and, as such, the Examiner's rejection should be withdrawn.

In the present case, the Examiner contended that the degree of amino acid identity among IMP and other integral membrane family proteins is insufficient to establish that IMP is a member of the integral membrane family of proteins and thus shares the same utilities. The Examiner attempted to support this assertion with the teachings of Bowie et al. (Science (1990) 247:1306-1310), and Burgess et al. (J. Cell Biol. (1990) 111:2129-2138), all of record and addressed below. However, all of these references fail to support the outstanding rejections.

Applicants submit that the teachings of Bowie et al. are, in part, counter to the outstanding rejections, and in part, supportive of the asserted utilities of IMP based on amino acid sequence homology to integral membrane family proteins. Careful review of this reference reveals that the teachings of Bowie et al. are directed primarily toward studying the effects of site-directed substitution of amino acid residues in certain proteins in order to determine the relative importance of these residues

to protein structure and function. As discussed below in further detail, such experiments are not relevant to Applicants' use of amino acid sequence homology to reasonably predict protein function.

In support of Applicants' use of amino acid sequence homology to reasonably predict the utility of the claimed polypeptide, Bowie et al. teach that evaluating sets of related sequences, which are members of the same gene family, is an accepted method of identifying functionally important residues that have been conserved over the course of evolution. (Bowie et al., p. 1306, 1st column, last paragraph, and 2nd column, 2nd full paragraph; p. 1308, 1st column, last paragraph; p. 1310, 1st column, last paragraph.) It is known in the art that natural selection acts to conserve protein function. As the Examiner stated and as taught by Bowie et al., proteins are tolerant of numerous amino acid substitutions that maintain protein function, and it is natural selection that permits these substitutions to occur. Conversely, mutations that reduce or abolish protein function are eliminated by natural selection. Based on these central tenets of molecular evolution, Applicants submit that the amino acid differences among Applicants' claimed polypeptide and known ion channel regulators are likely to occur at positions of minimal functional importance, while residues that are conserved are likely those that are important for protein function. One of ordinary skill in the art would further conclude that the level of conservation observed between Applicants' claimed polypeptide and ion channel regulators is indicative of a common function, and hence, common utility, among these proteins.

The Examiner further cited Burgess et al. as demonstrating the "sensitivity of proteins to alterations of even a single amino acid..." (Office Action at p. 5). However, these references are not relevant to the case at hand. Burgess et al. describe mutagenesis of HBGF-1 at an amino acid residue known to be important for ligand binding. In this cases, particular amino acid residues with known importance to protein function were specifically targeted for site-directed mutagenesis. These mutations were "artificially" created in the laboratory and, therefore, are **not** analogous to molecular evolution, which is profoundly influenced by natural selection. For example, the deactivating mutations as described by Burgess et al. would almost certainly not be tolerated in nature. Furthermore, it is clear that over the course of evolution, amino acid residues that are critical for protein function are **conserved**. Thus, the amino acid differences between SEQ ID NO:1 and stomatin and stomatin-like

proteins are likely to represent substitutions that do **not** alter protein function. Therefore, the teachings of Burgess et al. are not relevant to the case at hand.

One could then argue that partial loss-of-function mutations do occur in nature, for example, the mutation in hemoglobin that causes sickle cell anemia. However, this example is the **rare** exception in evolution, **not the rule**. Persistence of such a mutation in a population would **not** be expected by one of ordinary skill in the art. Persistence occurs only because of the fluke of heterozygous advantage. Therefore, the Examiner's assertion that one of skill in the art would routinely expect to find single amino acid substitutions that drastically affect the function of the individual members of a conserved protein family is entirely unsubstantiated.

IV. By Requiring the Patent Applicant to Assert a Particular or Unique Utility, the Patent Examination Utility Guidelines and Training Materials Applied by the Patent Examiner Misstate the Law

There is an additional, independent reason to withdraw the rejections: to the extent the rejections are based on Revised Interim Utility Examination Guidelines (64 FR 71427, December 21, 1999), the final Utility Examination Guidelines (66 FR 1092, January 5, 2001) and/or the Revised Interim Utility Guidelines Training Materials (USPTO Website www.uspto.gov, March 1, 2000), the Guidelines and Training Materials are themselves inconsistent with the law.

The Training Materials, which direct the Examiners regarding how to apply the Utility Guidelines, address the issue of specificity with reference to two kinds of asserted utilities: "specific" utilities, which meet the statutory requirements, and "general" utilities, which do not. The Training Materials define a "specific utility" as follows:

A [specific utility] is *specific* to the subject matter claimed. This contrasts to *general* utility that would be applicable to the broad class of invention. For example, a claim to a polynucleotide whose use is disclosed simply as "gene probe" or "chromosome marker" would not be considered to be specific in the absence of a disclosure of a specific DNA target. Similarly, a general statement of diagnostic utility, such as diagnosing an unspecified disease, would ordinarily be insufficient absent a disclosure of what condition can be diagnosed.

The Training Materials distinguish between “specific” and “general” utilities by assessing whether the asserted utility is sufficiently “particular,” *i.e.*, unique (Training Materials at p.52) as compared to the “broad class of invention.” (In this regard, the Training Materials appear to parallel the view set forth in Stephen G. Kunin, Written Description Guidelines and Utility Guidelines, 82 J.P.T.O.S. 77, 97 (Feb. 2000) (“With regard to the issue of specific utility the question to ask is whether or not a utility set forth in the specification is *particular* to the claimed invention.”).)

Such “unique” or “particular” utilities never have been required by the law. To meet the utility requirement, the invention need only be “practically useful,” *Natta*, 480 F.2d 1 at 1397, and confer a “specific benefit” on the public. *Brenner*, 383 U.S. at 534. Thus incredible “throwaway” utilities, such as trying to “patent a transgenic mouse by saying it makes great snake food,” do not meet this standard. Karen Hall, Genomic Warfare, *The American Lawyer* 68 (June 2000) (quoting John Doll, Chief of the Biotech Section of USPTO).

This does not preclude, however, a general utility, contrary to the statement in the Training Materials where “specific utility” is defined (p. 5). Practical real-world uses are not limited to uses that are unique to an invention. The law requires that the practical utility be “definite,” not particular. *Montedison*, 664 F.2d at 375. Applicants are not aware of any court that has rejected an assertion of utility on the grounds that it is not “particular” or “unique” to the specific invention. Where courts have found utility to be too “general,” it has been in those cases in which the asserted utility in the patent disclosure was not a practical use that conferred a specific benefit. That is, a person of ordinary skill in the art would have been left to guess as to how to benefit at all from the invention. In *Kirk*, for example, the CCPA held the assertion that a man-made steroid had “useful biological activity” was insufficient where there was no information in the specification as to how that biological activity could be practically used. *Kirk*, 376 F.2d at 941.

The fact that an invention can have a particular use does not provide a basis for requiring a particular use. See *Brana, supra* (disclosure describing a claimed antitumor compound as being homologous to an antitumor compound having activity against a “particular” type of cancer was determined to satisfy the specificity requirement). “Particularity” is not and never has been the *sine qua non* of utility; it is, at most, one of many factors to be considered.

As described *supra*, broad classes of inventions can satisfy the utility requirement so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. Only classes that encompass a significant portion of nonuseful members would fail to meet the utility requirement. *Supra* § III.B. (*Montedison*, 664 F.2d at 374-75).

The Training Materials fail to distinguish between broad classes that convey information of practical utility and those that do not, lumping all of them into the latter, unpatentable category of “general” utilities. As a result, the Training Materials paint with too broad a brush. Rigorously applied, they would render unpatentable whole categories of inventions heretofore considered to be patentable, and that have indisputably benefitted the public, including the claimed invention. *See supra* § III.B. Thus the Training Materials cannot be applied consistently with the law.

V. Summary of Arguments Regarding Utility Rejection

Applicants respectfully submit that rejections for lack of utility based, *inter alia*, on an allegation of lack of specificity as set forth in the Office Action and as justified in the Revised Interim and final Utility Guidelines and Training Materials, are not supported in the law. Neither are they scientifically correct, nor supported by any evidence or sound scientific reasoning. These rejections are alleged to be founded on facts in court cases such as *Brenner* and *Kirk*, yet those facts are clearly distinguishable from the facts of the instant application, and indeed most if not all nucleotide and protein sequence applications. Nevertheless, the PTO is attempting to mold the facts and holdings of these prior cases, “like a nose of wax,”² to target rejections of claims to polypeptide and polynucleotide sequences, as well as to claims to methods of detecting said polynucleotide sequences, where biological activity information has not been proven by laboratory experimentation, and they have done so by ignoring perfectly acceptable utilities fully disclosed in the specifications as well as well-established utilities known to those of skill in the art. As is disclosed in the specification, and even more clearly, as

²“The concept of patentable subject matter under §101 is not ‘like a nose of wax which may be turned and twisted in any direction * * *.’ *White v. Dunbar*, 119 U.S. 47, 51.” (*Parker v. Flook*, 198 USPQ 193 (US SupCt 1978))

one of ordinary skill in the art would understand, the claimed invention has well-established, specific, substantial and credible utilities. The rejections are, therefore, improper and should be withdrawn.

Moreover, to the extent the above rejections were based on the Revised Interim and final Examination Guidelines and Training Materials, those portions of the Guidelines and Training Materials that form the basis for the rejections should be determined to be inconsistent with the law.

Written description rejections under 35 U.S.C. § 112, first paragraph

Claims 1, 2, 17, and 18 were rejected under the first paragraph of 35 U.S.C. § 112 for an alleged lack of adequate written description. This rejection is traversed.

The requirements necessary to fulfill the written description requirement of 35 U.S.C. § 112, first paragraph, are well established by case law.

. . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession of *the invention*. The invention is, for purposes of the “written description” inquiry, *whatever is now claimed*. *Vas-Cath, Inc. v. Mahurkar*, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991).

Attention is also drawn to the Patent and Trademark Office’s own “Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1”, published January 5, 2001, which provide that:

An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics. What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail. If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met (*footnotes omitted*).

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:1 is specifically disclosed in the application (see, for example, the Sequence Listing at p. 53-54, and p.14, line 19 through p.15, line 11 of the Specification). Variants of SEQ ID NO:1 are described, for example, at p. 15, lines 17-20. In particular, the preferred, more preferred, and most preferred IMP variants (80%, 90%, and 95% amino acid sequence similarity to SEQ ID NO:1) are described, for example, at p. 15, lines 17-20. Incyte clones in which the nucleic acids encoding the human IMP were first identified and libraries from which those clones were isolated are described, for example, at p. 14, line 1. Chemical and structural features of SEQ ID NO:1 are described, for example, at p. 14, lines 7-18. SEQ ID NO:1, one of ordinary skill in the art would recognize naturally-occurring variants of SEQ ID NO:1 having 90% sequence identity to SEQ ID NO:1. Fragments of IMP could be made using either recombinant methods (e.g., see pp. 19-24) or by chemical synthesis (e.g., see p. 17, lines 15-22, and p. 24, lines 13-19). “Biologically-active” fragments of IMP are defined at p. 8, line 16-17. Methods for determining biological activity of IMP and fragments thereof are provided, e.g., at p. 51, line 22-26. The Specification at pp. 14-15 also discusses the biological activity of protein homologs of IMP. Accordingly, the Specification provides an adequate written description of the recited polypeptide sequences.

A. The Specification provides an adequate written description of the claimed “variants” of SEQ ID NO:1.

The Office Action has further asserted that the claims are not supported by an adequate written description because

“[t]he written description in this case only sets forth SEQ ID NO:1 and therefore the written description is not commensurate in scope with the claims which read on variants which as claimed, include naturally occurring polypeptides that are at least 90% identical to SEQ ID NO:1, and biologically active and immunogenic fragments of SEQ ID NO:1.” Office Action mailed January 28, 2003; at p. 8, lines 5-9.

Such a position is believed to present a misapplication of the law.

1. The present claims specifically define the claimed genus through the recitation of chemical structure

Court cases in which “DNA claims” have been at issue (which are hence relevant to claims to proteins encoded by the DNA and antibodies which specifically bind to the proteins) commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

In claims to genetic material, however, a generic statement such as “vertebrate insulin cDNA” or “mammalian insulin cDNA,” without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. §112; *i.e.*, “an mRNA of a vertebrate, which mRNA encodes insulin” in *Lilly*, and “DNA which codes for a human fibroblast interferon-beta polypeptide” in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polypeptides in terms of chemical structure, rather than on functional characteristics. For example, the “variant language” of independent claim 1 recites chemical structure to define the claimed genus:

2. An isolated and purified polynucleotide sequence encoding a polypeptide selected from the group consisting of:...b) a naturally-occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1...

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. In the present case, there is no reliance merely on a description of functional characteristics of the polypeptides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polypeptides. The polypeptides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry “on whatever is now claimed,” the Office Action failed to provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*.

2. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the “dark ages” of recombinant DNA technology.

The present application has a priority date of January 9, 1997. Much has happened in the development of recombinant DNA technology in the 17 or more years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polynucleotide variants at the time of filing of this application.

3. Summary

The Office Action failed to base its written description inquiry “on whatever is now claimed.” Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polypeptides defined by the present claims is adequately described, as evidenced by Brenner et al.

and consideration of the claims of the '740 patent involved in *Lilly*. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action.

CONCLUSION

Applicants respectfully submit that rejections for lack of utility based, *inter alia*, on an allegation of lack of specificity, as set forth in the Office Action and as justified in the Revised Interim and final Utility Guidelines and Training Materials, are not supported in the law. Neither are they scientifically correct, nor supported by any evidence or sound scientific reasoning. These rejections are alleged to be founded on facts in court cases such as *Brenner* and *Kirk*, yet those facts are clearly distinguishable from the facts of the instant application, and indeed most if not all nucleotide and protein sequence applications. Nevertheless, the PTO is attempting to mold the facts and holdings of these prior cases, "like a nose of wax,"³ to target rejections of claims to polypeptide and polynucleotide sequences, as well as to claims to methods of detecting said polynucleotide sequences, where biological activity information has not been proven by laboratory experimentation, and they have done so by ignoring perfectly acceptable utilities fully disclosed in the specifications as well as well-established utilities known to those of skill in the art. As is disclosed in the specification, and even more clearly, as one of ordinary skill in the art would understand, the claimed invention has well-established, specific, substantial and credible utilities. The rejections are, therefore, improper and should be withdrawn.

³"The concept of patentable subject matter under §101 is not 'like a nose of wax which may be turned and twisted in any direction * * *.' *White v. Dunbar*, 119 U.S. 47, 51." (*Parker v. Flook*, 198 USPQ 193 (US SupCt 1978))

In light of the above remarks, Applicants submit that the present Application is fully in condition for allowance, and request that the Examiner withdraw the outstanding rejections. Early notice to that effect is earnestly solicited.

If the Examiner contemplates other action, or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact Applicants' Agent at (650) 845-5415.

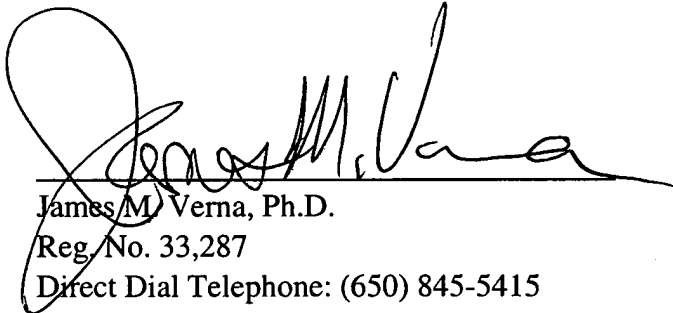
Applicants believe that no fee is due with this communication. However, if the USPTO determines that a fee is due, the Commissioner is hereby authorized to charge Deposit Account No. 09-0108.

Respectfully submitted,

INCYTE CORPORATION

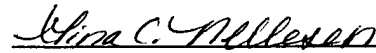
Date:

November 14, 2003


James M. Verna, Ph.D.
Reg. No. 33,287
Direct Dial Telephone: (650) 845-5415

Date:

November 14, 2003


Gina C. Nellesen
Reg. No. 52,062
Direct Dial Telephone: (650) 843-7342

Customer No.: 27904
3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

Enclosures:

1. Wang et al., Identification and characterization of human SLP-2, a novel homologue of stomatin (Band 7.2b) present in erythrocytes and other tissues, Journal of Biological Chemistry 2000;275(11):8062-8071.
2. Alignment 789094CD1 v. AF190167
3. Owczarek C.M. et al., A novel member of the STOMATIN/EPB72/mec-2 family, stomatin-like 2 (STOML2), is ubiquitously expressed and localized to HSA chromosome 9p13.1, Cytogenet Cell Genet. 2001;92(3-4):196-203.
4. Alignment 789094CD1 v. AF282596
5. Fricke et al., Stomatin immunoreactivity in ciliated cells of the human airway epithelium. Anat Embryol (Berl). 2003 Jul;207(1):1-7, Epub 2003 May 21.
6. Brenner et al., Proc. Natl. Acad. Sci. U.S.A. 95:6073-78 (1998).
7. John C. Rockett, et. al., Differential gene expression in drug metabolism and toxicology: practicalities, problems, and potential, Xenobiotica 29:655-691 (July 1999).
8. Emile F. Nuwaysir, et al., Microarrays and Toxicology: The Advent of Toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999).
9. Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000).
10. Email from the primary investigator, Dr. Cynthia Afshari to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding.

Identification and Characterization of Human SLP-2, a Novel Homologue of Stomatin (Band 7.2b) Present in Erythrocytes and Other Tissues*

(Received for publication, October 19, 1999, and in revised form, December 18, 1999)

Yingjian Wang and Jon S. Morrow†

From the Department of Pathology Yale University School of Medicine, New Haven, Connecticut 06510

Human stomatin (band 7.2b) is a 31-kDa erythrocyte membrane protein of unknown function but implicated in the control of ion channel permeability, mechanoreception, and lipid domain organization. Although absent in erythrocytes from patients with hereditary stomatocytosis, stomatin is not linked to this disorder. A second stomatin homologue, termed SLP-1, has been identified in nonerythroid tissues, and other stomatin related proteins are found in *Drosophila*, *Caenorhabditis elegans*, and plants. We now report the cloning and characterization of a new and unusual stomatin homologue, human SLP-2 (stomatin-like protein 2). SLP-2 is encoded by an ~1.5-kilobase mRNA (GenBank™ accession no. AF190167). The gene for human SLP-2, *HUSLP2*, is present on chromosome 9p13. Its derived amino acid sequence predicts a 38,537-kDa protein that is overall ~20% similar to human stomatin. Northern and Western blots for SLP-1 and SLP-2 reveal a wide but incompletely overlapping tissue distribution. Unlike SLP-1, SLP-2 is also present in mature human erythrocytes (~4,000 ± 5,600 (± 2 S.D.) copies/cell). SLP-2 lacks a characteristic NH₂-terminal hydrophobic domain found in other stomatin homologues and (unlike stomatin) is fully extractable from erythrocyte membranes by NaOH, pH 11. SLP-2 partitions into both Triton X-100-soluble and -insoluble pools in erythrocyte ghost membranes or when expressed in cultured COS cells and migrates anomalously on SDS-polyacrylamide gel electrophoresis analysis with apparent mobilities of ~45,500, 44,600, and 34,300 M_r. The smallest of these protein bands is believed to represent the product of alternative translation initiated at AUGs beginning with nt 217 or 391, although this point has not been rigorously proven. Collectively, these findings identify a novel and unusual member of the stomatin gene superfamily that interacts with the peripheral erythrocyte cytoskeleton and presumably other integral membrane proteins but not directly with the membrane bilayer. We hypothesize that SLP-2 may link stomatin or other integral membrane proteins to the peripheral cytoskeleton and thereby play a role in regulating ion channel conductances or the organization of sphingolipid and cholesterol-rich lipid rafts.

Previously known as band 7.2b because of its relative electrophoretic mobility in samples of human red blood cell ghost preparations, stomatin is a less characterized integral erythrocyte membrane protein with a molecular mass of 31 kDa (1, 2). Deficiency of stomatin in red cells is associated with hereditary stomatocytosis, a disease with marked red cell shape abnormalities and increased monovalent cation permeability (for reviews, see Refs. 3 and 4). However, linkage studies and direct sequencing establish that a defect in stomatin is not the cause of this disorder (4–6), and mice lacking stomatin retain normal red cell morphology and apparently normal function (7). The role of stomatin thus remains a mystery. In a human amniotic cell line, stomatin concentrates preferentially in plasma membrane protrusions and appears to co-localize with cortical actin microfilaments (8). In *Caenorhabditis elegans*, a stomatin homologue (MEC-2) is required for sensory mechanoreception and the gating of an oligomeric sodium channel (9). A second homologue in *C. elegans* (UNC-24), a protein most similar to a human stomatin homologue termed SLP-1 (10), is required for normal locomotor response to volatile anesthetics and contains a region of sequence homologous to the nonspecific lipid transfer protein (11). A third homologue in *C. elegans* (UNC-1) also appears to play a central role in the organism's response to volatile anesthetics (12). In plants, a homologue of stomatin (*slp*) is required for bean nodulation and growth in media containing hypertonic monovalent cations (13). Together, these data implicate stomatin (or a homologue) as an adapter between ion channels and the cytoskeletal network, perhaps influencing channel stability and organization in the plasma membrane. Other observations suggest that stomatin binds calpromotin (involved with the activation of the charybdotoxin-sensitive calcium-dependent potassium channel of red cells) (14) and participates in the trans-bilayer exchange or reorientation of phospholipids (15, 16).

The structure or disposition of stomatin in the membrane is not well defined; available data suggest an unusual topography. Sequence analysis predicts that stomatin has a single 23-residue hydrophobic domain near its NH₂ terminus, and it is palmitoylated just proximal to this predicted hydrophobic domain on Cys²⁹ (in the mature protein, Cys³⁰ in the derived sequence (17)). Sequences distal to the putative transmembrane segment are hydrophilic and are predicted to form a bipartite β-sheet and α-helical structure (3). The large COOH-terminal domain is cytoplasmic, based on its selective protease sensitivity in leaky ghosts but not intact red cells (18). The protein is phosphorylated on Ser⁹, and the short NH₂-terminal sequence containing the phosphorylation site (which is proximal to the hydrophobic domain) is also cytoplasmic in its orientation (19). Thus, both ends of the protein must face the cytosol. It is unknown whether other portions of stomatin, such as the predicted β-sheet region, enter the bilayer or whether the hydrophobic region of stomatin enters the bilayer but does

* This work was supported by National Research Award F32-HL09977 (to Y. W.) and by grants from the National Institutes of Health (to J. S. M.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Portions of this work have been presented in abstract form (41).

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AF190167.

† To whom correspondence should be addressed. Tel.: 203-785-3624; Fax: 203-785-7037; E-mail: jon.morrow@yale.edu.

not span it. In the membrane, stomatin appears as large ($n \approx 9-12$) homo-oligomers, a property bestowed by sequences near the COOH terminus (20).

The identification of stomatin homologues has provided important insights into the potential functions of this gene superfamily. Three homologues have been identified in *C. elegans*: MEC-2, UNC-24, and UNC-1. As noted above, MEC-2 appears to link degenerin channels, homologues of mammalian epithelial sodium channels, to a microtubule based cytoskeletal network; UNC-24 and UNC-1 bestow sensitivity to certain volatile anesthetics. The only vertebrate homologue of stomatin identified is human SLP-1, which is most abundant in the brain and shares many similarities with UNC-24 (10). All of these proteins as well as the stomatin from other species (e.g. mouse and zebra fish) share a characteristic NH_2 -terminal hydrophobic domain as well as a consensus stomatin signature sequence that defines the stomatin gene family (i.e. $\text{RX}_2(\text{L/I/V})(\text{S/A/N})\text{X}_6(\text{L/I/V})\text{DX}_2\text{TX}_2\text{WG}(\text{L/I/V})(\text{K/R/H})(\text{L/I/V})\text{X}(\text{K/R})(\text{L/I/V})\text{E}(\text{L/I/V})(\text{K/R})$ (as defined by the PROSITE program, using data derived from Ref. 21).

We now report the cloning and characterization of a new member of the vertebrate stomatin gene family. We name this gene, identified in a human heart cDNA library, *HUSLP2*, and the derived protein SLP-2 (stomatin-like protein 2). Similar to other family members, SLP-2 shares the cognate stomatin signature sequence noted above. However, it is the first member of this family to be recognized that lacks a NH_2 -terminal hydrophobic domain. SLP-2 is widely expressed in many tissues, as is SLP-1, but unlike SLP-1 it is also found in the mature human erythrocyte membrane. In the erythrocyte, it associates with the cortical spectrin-actin cytoskeleton and probably with other integral membrane proteins but is not itself integral to the membrane bilayer. In the erythrocyte membrane, it also appears to exist at least partially as an oligomeric protein complex. These features distinguish it from stomatin and SLP-1 and suggest that members of this gene superfamily may function as both integral and peripheral membrane proteins. The identification of SLP-2 as a second stomatin-related protein in red cells and as a new component of the peripheral membrane skeleton also may have implications for understanding the phenotype of certain hemolytic conditions.

MATERIALS AND METHODS

Cloning and Sequencing—Unless otherwise stated, all molecular biological procedures followed standard methods (22). Candidate sequences were amplified from a Marathon-Ready cDNA library prepared from human heart muscle (CLONTECH, Palo Alto, CA). The Advantage cDNA PCR¹ kit was used to perform 5'- and 3'-RACE amplifications, following the instructions of the manufacturer (CLONTECH). For 5'-RACE, the primer used was GTCCCCAGACTCCCTGGCC; the primer for the 3'-RACE was GGCGGTTGGGAATGCTGGCGG. PCR products were purified by agarose gel electrophoresis and cloned into TA cloning vector (Invitrogen). All constructs were amplified, cloned, and sequenced multiple times to verify the fidelity of the cDNA sequences obtained. Automated DNA sequencing was carried out by the Keck Laboratory (Yale University). FLAG-tagged SLP-2 constructs were prepared using a synthetic oligonucleotide representing the coding sequence of the last seven amino acids of stomatin followed by the FLAG sequence and a stop codon (23), paired with a primer corresponding to the desired ATG initiator codon. The PCR product was cloned into the pSG5 expression vector (Stratagene) prior to transfection into either COS or 293T cells.

Northern Blot—Northern blot analysis of multiple human tissues was performed according to the instructions of the manufacturer, using their multiple human tissue Northern blot (catalog no. 7760, lot

TABLE I
List of EST stomatin-like sequences and their genes

Number	Origin	Genes
emb-aa351669	Human	SLP-1
gb-aa364511	Human	SLP-1
gb-H05582	Human	SLP-1
gb-h06392	Human	SLP-1
gb-t77395	Human	SLP-1
gb-R12214	Human	SLP-1
gb-aa072966	Mouse	SLP-1
gb-W81294	Human	HUSLP2
gb-aa910939	Human	HUSLP2
gb-A1240857	Human	HUSLP2
gb-R81533	Human	HUSLP2
gb-A116846	Human	MUSLP2
gb-aa270458	Mouse	MUSLP2
gb-al313898	Mouse	MUSLP2

7010558 (CLONTECH)). Randomly ³²P-labeled *Bam*HI fragments of SLP-1 or SLP-2 cDNA were used as hybridization probes. The loading of mRNA was verified by probing β -actin mRNA with the probe provided in the multiple human tissue Northern blot kit.

Antibody Production—Antibodies were raised in New Zealand White rabbits as before (24). The cDNAs of SLP-1 and SLP-2 in pCR2.1 (Invitrogen) were digested with *Bam*HI and *Eco*RI, and the *Bam*HI/*Eco*RI fragments were cloned into pGEX-3X (Amersham Pharmacia Biotech) to produce recombinant GST fusion proteins representing the approximate C-terminal two-thirds of each protein (corresponding to amino acids 65–290 for SLP-2). These were expressed in the DH5 α strain of *Escherichia coli* to generate the corresponding recombinant fusion proteins. The fusion proteins were analyzed by SDS-PAGE and stained lightly with Coomassie Blue, and the recombinant proteins were sliced from the gel, emulsified in incomplete Freund's adjuvant, and used as antigens for immunization. Sera were affinity-purified, and the reactivity to GST was removed by immunosorption with GST immobilized on agarose beads. Antibodies directed against a mouse full-length stomatin GST fusion protein prepared in a similar fashion were kindly provided by Paul Stabach and Dr. John Sinard.²

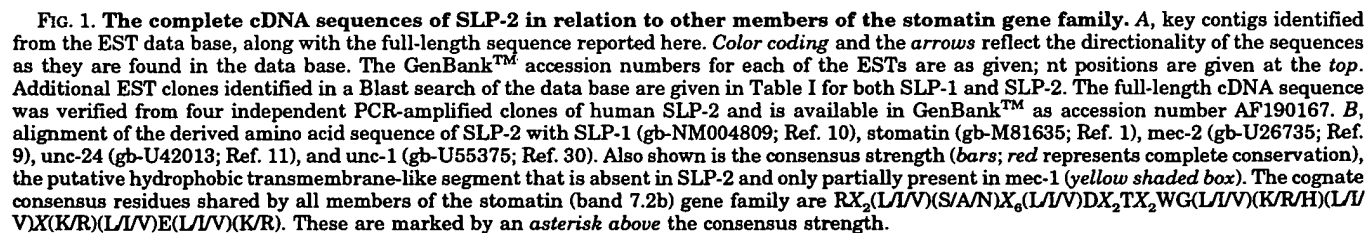
Immunofluorescence—Fresh human erythrocytes were washed twice with chilled PBS and then air-dried as smears on glass slides. These were then fixed in methanol on ice for 10 min. The cells were then washed with cold PBS and blocked with 5% BSA (w/v) in PBS for 1 h. Primary antibodies were applied overnight in a humidified chamber, followed by PBS rinse. Cy3-labeled secondary anti-rabbit antibodies were then applied for 2 h. After washing, slides were mounted with glass coverslips and viewed by epifluorescence or by confocal microscopy using an Olympus AX-70 inverted confocal microscope.

Cell Preparations and Extraction—Cell lines were from ATCC and were transfected using LipofectAMINETM (Stratagene), following the manufacturer's protocol. Fresh human erythrocytes and erythrocyte ghosts were prepared by washing twice with cold PBS, followed by lysis in a 20-fold volume of 5 mM sodium phosphate, pH 7.5, in the presence of 1 mM EDTA and various protease inhibitors (25). Triton extraction was carried out at 4 °C for 15 min by suspending approximately 1×10^6 cells (erythrocyte ghosts or cultured cells) in 1% Triton in PBS. Triton-soluble and -insoluble fractions were separated by centrifugation for 30 min at $\sim 30,000 \times g$. Packed ghosts were alkaline extracted by incubation in $\sim 10 \times$ volume of 15 mM NaOH for 15 min at 4 °C; extractable and inextractable fractions were separated by centrifugation as above. Salt extractions were carried out by first incubating freshly prepared erythrocyte ghosts with 0.1 mM EDTA at pH 8–9 at 37 °C for 30 min. The pellet resulting from this extraction (largely consisting of erythrocyte inside-out vesicles) was then incubated with 0.5 M KCl under the same conditions for an additional 30 min.

Other Procedures—For Western blotting, cells or tissues were lysed in lysis buffer (2% SDS in PBS plus protease inhibitors) and separated by SDS-PAGE. After transfer to polyvinylidene difluoride membrane, proteins of interest were detected with affinity-purified antibodies. *In vitro* translations were performed with the TNTTM coupled reticulocyte lysate system (Promega) following the manufacturer's instructions. Each reaction used 2 μ g of plasmid DNA. SDS-PAGE analysis followed the method of Laemmli (26). Protein determinations were carried out using the Pierce BCA method (product 23225), as described in Ref. 27.

¹ The abbreviations used are: PCR, polymerase chain reaction; RACE, rapid amplification of cDNA ends; PAGE, polyacrylamide gel electrophoresis; GST, glutathione S-transferase; PBS, phosphate-buffered saline; nt, nucleotide; EST, expressed sequence tag.

² P. Stabach and J. Sinard, unpublished observations.



B(continued)

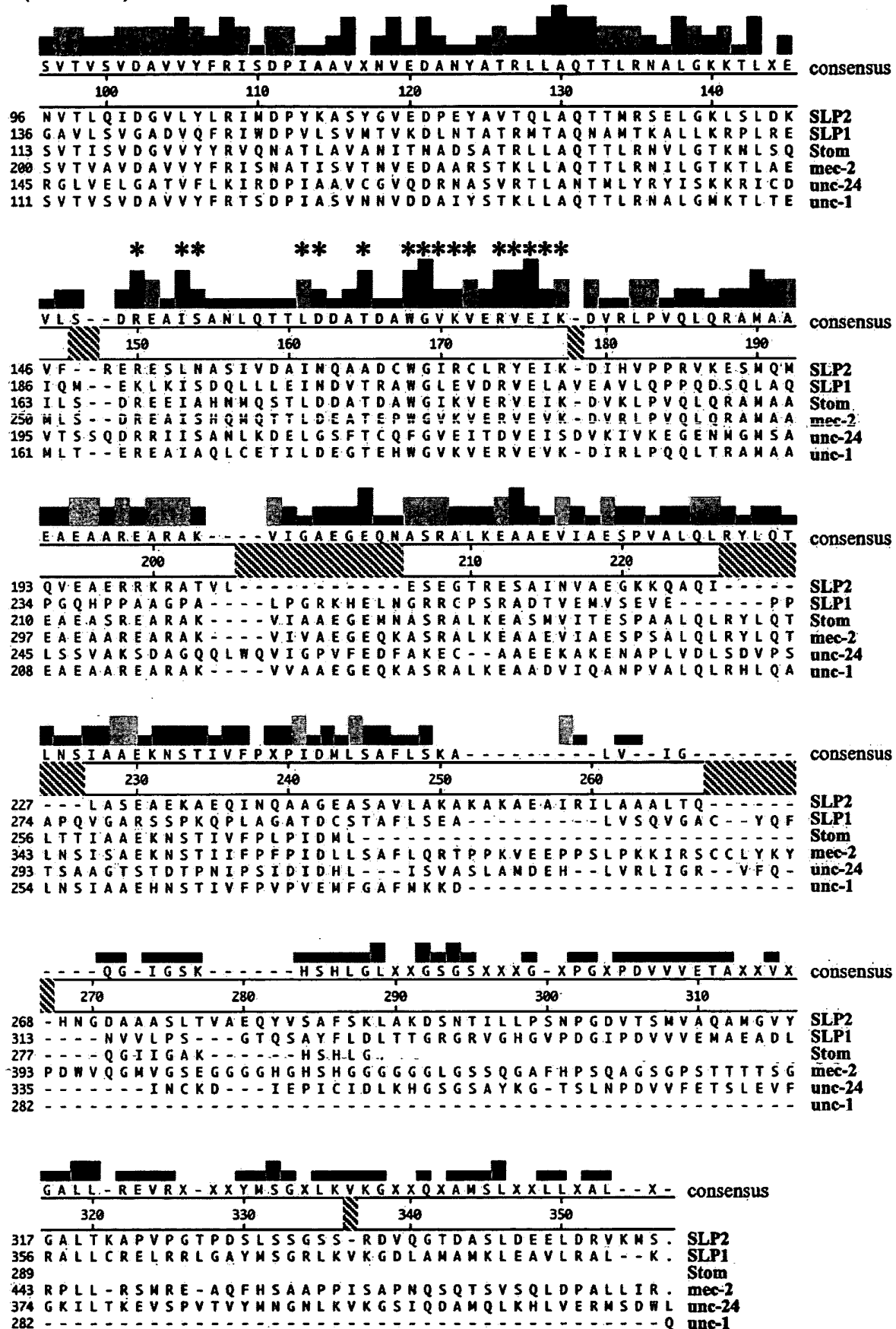


FIG. 1—continued

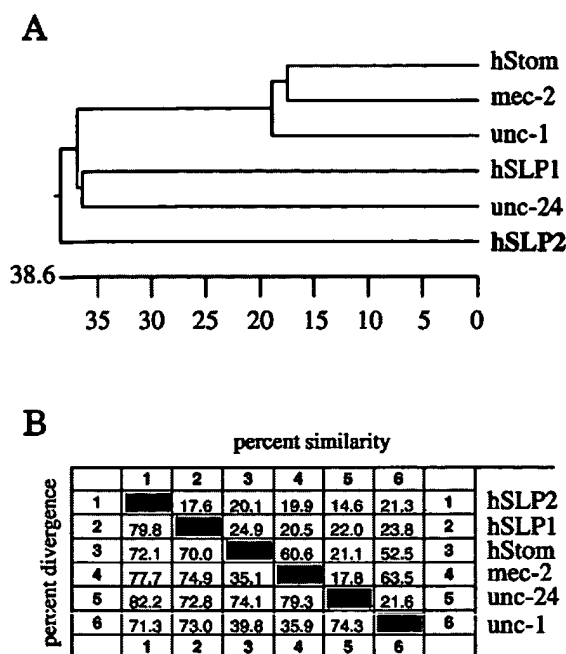


FIG. 2. The relatedness of different stomatins. A, dendrogram of the relationships between members of the stomatin superfamily. Dendrograms were constructed using the CLUSTAL procedure (39), which allows estimation of relatedness and provides an approximation to an evolutionary tree. Since this program constructs the dendrogram on the basis of pairwise matches and the formation of consensus sequences, it does not technically provide an evolutionary tree. It results in a dendrogram that shows relatedness as a function of length of each of the branches, where the length is proportional to sequence distance. Note that SLP-2 defines a new branch of the stomatin gene superfamily. B, overall level of sequence homology and divergence within the stomatin gene family. The similarities and divergences shown are over the full-length sequences. When only the consensus region that is distal to the hydrophobic COOH-terminal domain is considered, the degree of similarity between these proteins rises (see Fig. 3).

RESULTS

Cloning of cDNA Encoding SLP-2—A BLAST search (28) of the EST data base using human stomatin cDNA sequence (1) identified several human and mouse sequences that encoded proteins with significant homology to stomatin. Sequence analysis revealed that they derived from two distinct genes. The most significant of the dbEST sequences are listed in Table I. One set of sequences belonged to human SLP-1. The others identified a novel gene, SLP-2, as described in this report. These sequences were used to select oligonucleotide primers that enabled the PCR amplification of the relevant sequences. To extend the sequence of SLP-2 at both 5'- and 3'-ends, we performed RACE by PCR using a Marathon-Ready cDNA library from human heart (CLONTECH, Palo Alto). Several overlapping clones were sequenced, and a 1,303-bp cDNA that included the full-length sequence of SLP-2 was assembled (Fig. 1A and GenBank™ accession no. AF190167). Sequence analysis revealed a coding region of 1071 nt, predicting a 357-amino acid, 38,537-kDa protein (Fig. 1B). The sequence at the 5'-end of the mRNA is interesting for the presence of three potential ATG initiator sites, all sharing the same open reading frame, as discussed below. The presumptive start site begins at nt 64. A satisfactory Kozak initiation sequence is present immediately upstream of this ATG (29), but since the sequence upstream of this ATG is relatively short and in frame, there was concern whether all coding sequence had been identified. To address this issue, 12 additional clones were sequenced from a series of four additional PCRs, and the genomic data base was exhaustively searched for evidence of ESTs that extended the

SLP-2 sequence further upstream. In no case was additional 5'-sequence identified. From the human genome data base, SLP-2 was identified on chromosome 9p13 (GenBank™ accession no. AC004472). Analysis of this sequence using the program GeneTool™ (BioTools, Inc.), which is designed to find potential exons in genomic sequences, satisfactorily identified most of the exons responsible for the expressed SLP-2 sequence but did not predict any convincing exons upstream of the ATG beginning at nt 64. *In vivo* and *in vitro* expression of the SLP-2 cDNA also generated proteins of the correct size (see below). Based on these criteria, it was concluded that the full-length SLP-2 is as shown in Fig. 1.

SLP-2 Lacks the Hydrophobic Domain Found in Other Family Members—The derived amino acid sequence of SLP-2 is compared in Fig. 1B with the sequences of human SLP-1 (GenBank™ accession no. NM004809; Ref. 10); human stomatin (GenBank™ accession no. M81635; Ref. 1); mec-2 (GenBank™ accession no. U26735; Ref. 9); unc-24 (GenBank™ accession no. U42013; Ref. 11), and unc-1 (GenBank™ accession no. U55375; Ref. 30). All members of this family including SLP-2 share the cognate consensus sequence $RX_2(L/I/V)(S/A/N)X_6(L/I/V)DX_2TX_2WG(L/I/V)(K/R/H)(L/I/V)X(K/R)(L/I/V)E(L/I/V)(K/R)$ that defines proteins of this gene superfamily. Clustal analysis revealed that SLP-2 defines a new branch of the superfamily (Fig. 2A), approximately equidistant between stomatin and SLP-1 (Fig. 2B). However, when the predicted secondary structure and hydrophobicity of SLP-2 is compared with other stomatin family members, significant differences are apparent (Fig. 3). All previously recognized stomatin family members share a characteristic NH_2 -terminal hydrophobic domain, and most have a consensus sequence for palmitoylation centered on Cys²⁹ (17). These properties presumably allow stomatin, SLP-1, and close homologues to intercalate directly into the lipid bilayer. Neither of these features (*i.e.* a site for palmitoylation or a hydrophobic domain) are present in SLP-2 (Fig. 3). However, immediately distal to the missing hydrophobic domain (TM in Fig. 3), SLP-2 shares strong sequence homology with both stomatin and SLP-1 in the region predicted to contain β -sheet and α -helix structure. The overall amino acid composition of SLP-2 is also similar to other stomatins, although it is predicted to be a bit more basic with an anticipated isoelectric point of 7.3. Other predicted biophysical properties, and its composition are shown in Fig. 4.

SLP-2 Is Present in Most Tissues, Including Mature Human Erythrocytes—To examine the tissue distributions of the major stomatin gene superfamily members, antibodies were prepared to recombinant human stomatin, human SLP-1, and human SLP-2. After affinity purification and absorption against GST, these antibodies distinguished stomatin from SLP-1 and SLP-2 with high fidelity and little residual activity against GST (Fig. 5). These antibodies and the SLP-2 and SLP-1 cDNAs were then used to examine the distribution of these proteins in a variety of tissues by Western and Northern blotting (Fig. 6). Previous studies have established the distribution of stomatin (1, 2). Although SLP-2 was initially derived from a heart library, its ~1.5-kilobase mRNA was readily detected in all tissues examined (Fig. 6A). These included heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas. When normalized to the corresponding actin mRNA signal, the tissues with the highest SLP-2 mRNA levels were heart, liver, and pancreas. This distribution was distinct from the pattern of SLP-1 expression, in which the highest mRNA levels were found in brain and heart, with lesser but still detectable amounts in the other tissues.

A range of tissues and cell lines were examined with the antibodies to SLP-1 and SLP-2. The SLP-1 antibody did not

FIG. 3. SLP-2 lacks the amino-terminal hydrophobic domain of the other stomatin family members. The predicted secondary structure of SLP-2, SLP-1, and stomatin is shown based on the Chou and Fasman algorithm as implemented in the program ProteanTM (DNASar, Inc.) (40). Also shown is the hydrophobicity profile for each protein, and the putative transmembrane hydrophobic region of stomatin and SLP-1 is shaded yellow. Note that SLP-2 lacks this region, but beyond this region it shares other features characteristic of stomatin and SLP-1, including a proximal region predicted to be rich in β structure and a more distal region rich in α -helix. The overall similarity of SLP-2 and SLP-1 to stomatin over this region is given, for consensus lengths of 160 and 201 residues, respectively. A, α -helix; B, β -sheet; T, turns.

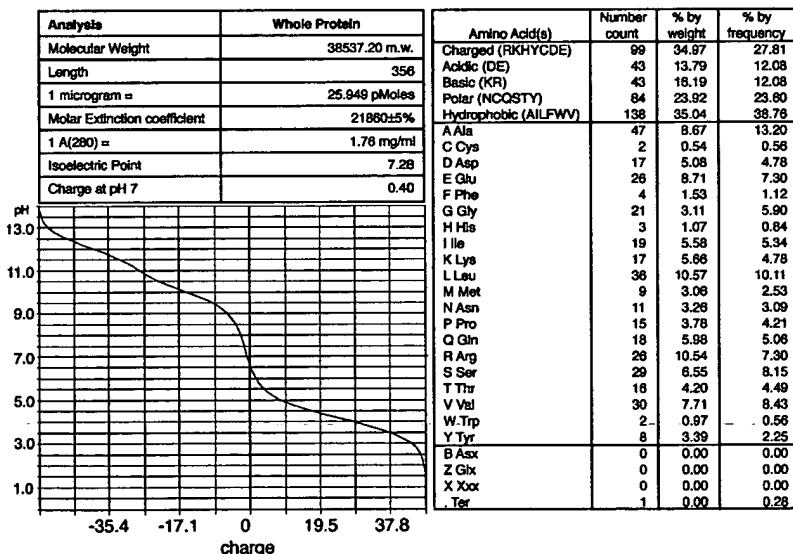
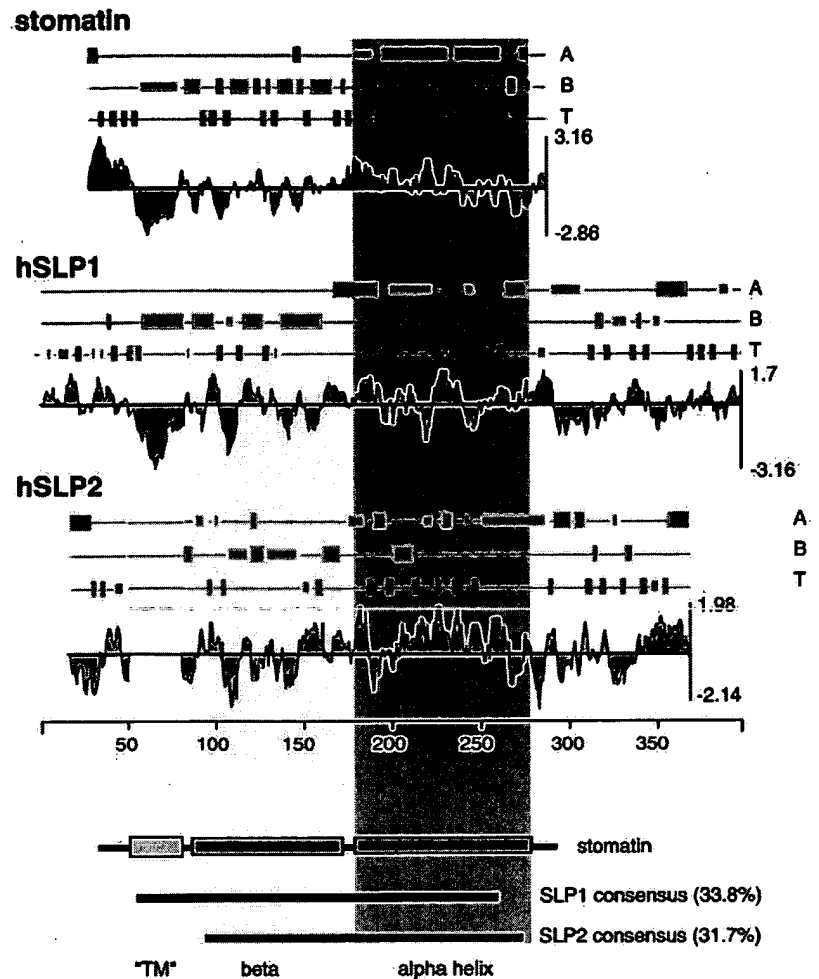


FIG. 4. Composition and predicted biophysical features of the SLP-2 protein. The protein is predicted to have an isoelectric point of 7.3 and its composition is conventional. These analyses were carried out using the program ProteanTM.

detect any protein in these blots; the reasons for this are unknown but may relate to the apparent reduced sensitivity of this antibody compared with the antibody against SLP-2 (data not shown). The antibodies to SLP-2 revealed a pattern that correlated well with its mRNA expression profile (Fig. 6B). In most tissues, the SLP-2 antibody detected a band at either M_r ~45,500 or ~44,600. Both of these bands are substantially larger than the predicted size of SLP-2 (38,537 kDa). In COS cells, A431 cells, and red blood cells, both the M_r 45,500 and

44,600 bands were evident, although the larger band was most prominent. These three cell types also displayed a faint immunoreactive band at M_r ~34,300. The origin of the multiple bands is unknown (see below). SLP-2 thus represents a novel stomatin gene superfamily member, one with an unusual structure. It is also a previously unrecognized component of the mature erythrocyte membrane.

Recombinant SLP-2 Migrates as Multiple Bands Comparable with Those of Wild Type SLP-2—The calculated size of SLP-2 is

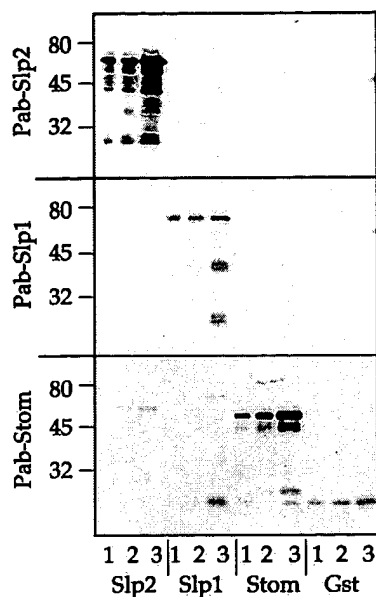


FIG. 5. Affinity-purified polyclonal antibodies discriminate between human SLP-2, SLP-1, and stomatin. Total bacterial lysates containing GST fusion proteins of SLP-2, SLP-1, stomatin (*Stom*), or GST alone were separated by SDS/PAGE and Western blotted using affinity-purified rabbit polyclonal antibodies against each of these proteins. The proteins loaded in each lane are as designated. The amounts of protein loaded into each lane were as follows: 10 ng (lane 1), 20 ng (lane 2), and 40 ng (lane 3). The antibodies used for blotting each gel, along with the M_r of three molecular weight standards, is shown to the left of each gel.

38,537 kDa, substantially smaller than the major bands observed in the tissue Western blots. These observations again raised the question of whether the cDNA that had been characterized represented the full-length product or whether the additional bands arose from post-translational modification, proteolysis, alternative initiation, alternative mRNA splicing, or protein associations not dissociated under the conditions of SDS-PAGE. To address these questions, SLP-2 was expressed either in cell-free lysates or after transfection into COS cells (Fig. 7). In cell-free rabbit reticulocyte lysates, full-length SLP-2 cDNA in the pCR2.1 vector generated a M_r ~45,500 protein as its major product, along with a doublet at ~34,500 and a smaller product with M_r ~26,300. When a second cDNA was used that contained a stop codon downstream of the initiator ATG (the codon beginning at nt 64), only the M_r ~34,500 product was generated as a major band, with two faint bands detected near M_r ~40,000. SLP-2 was also expressed in cultured COS and 293T cells, where the Triton X-100 solubility of the transfected products could be evaluated along with their apparent molecular weights. In these experiments, the eight-residue FLAG epitope tag was incorporated onto the COOH terminus of SLP-2 (23). Four SLP-2 bands were evident in the soluble fraction of COS cells: M_r ~45,500, ~44,600, ~34,300, and ~26,300 (Fig. 7B, lane 1). Although differing in their relative abundance, this pattern is similar to that observed for wild type SLP-2 in COS cells (cf. Fig. 6). Only the largest two bands (M_r 45,500 and 44,600) were present in the detergent-insoluble pellet and together constituted about 60% of the total FLAG-SLP-2 in these cells (Fig. 7B, lane 2). When the same full-length construct was expressed in 293T cells, the predominant product generated was at ~44,600, with only a small amount of the M_r 45,500 product evident. Both of these proteins were fully soluble after detergent extraction (Fig. 7B, lanes 5 and 6). Finally, a FLAG-tagged truncated SLP-2 construct was prepared that deleted the first initiator ATG (Fig. 7B, lanes 3 and 4). Expression of this construct yielded a single

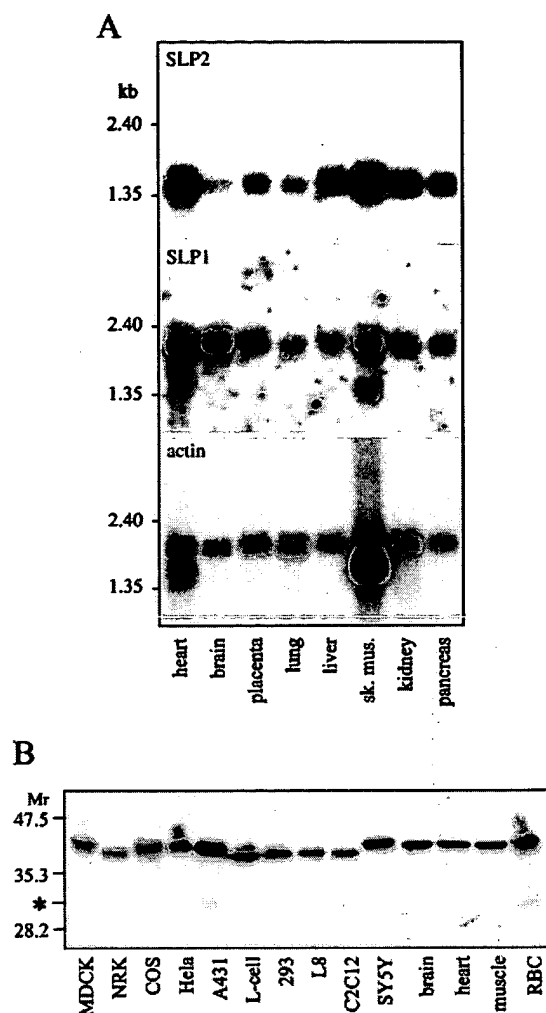


FIG. 6. Stomatatin-like proteins (SLP-1 and SLP-2) are widely distributed. A, Northern blot of electrophoretically separated human poly(A)⁺ mRNAs (~2 mg) from various adult tissues probed for SLP-1 and SLP-2. The same blot was hybridized for actin as a control for RNA loading. The positions of standards at 1.35 and 2.4 kilobases are as indicated. The SLP-2 mRNA is ~1.5 kilobases. B, Western blot demonstrating SLP-2 in various cultured cell lines and human tissues. Similar experiments carried out with the available antibodies for SLP-1 did not detect SLP-1 in these blots. Note the presence of strongly SLP-2-immunoreactive bands at M_r ~45,500 and ~44,600. Most cells and tissues contained only one of these two bands, although both were present in COS cells, A431 cells, and red blood cell ghosts. Also detected in Madin-Darby canine kidney, COS, and A431 cells and in human heart, skeletal muscle, and red blood cells was a faint immunoreactive band at M_r ~34,300 (*). Additional minor bands can also be detected in more heavily loaded samples (e.g. see Fig. 8). The origins of the minor bands and the reasons for the anomalous migration of the major SLP-2 band on SDS-PAGE (calculated mass of 38,537 Da), are unknown (see "Discussion"). The apparent molecular weights (M_r) are shown along the ordinate for three standard proteins (ovalbumin, 47,500; carbonic anhydrase, 35,300; and soybean trypsin inhibitor, 28,200). All protein bands were visualized by ECL autofluorography.

soluble product at M_r ~40,000. The sequences flanking the three potential initiator sites in the *SLP-2* gene, along with the calculated M_r values of the resultant products, are given in Fig. 7C.

SLP-2 Associates with the Triton-insoluble Cytoskeleton in Erythrocyte Ghosts—Quantitation of the amount of SLP-2 immunoreactive material in erythrocytes by comparative Western blotting using recombinant SLP-2 as a standard revealed $4,000 \pm 5,600$ (\pm S.D.) copies/red cell (data not shown). The disposition of SLP-2 in mature human erythrocytes was examined based on its resistance to extraction by detergent or salt.

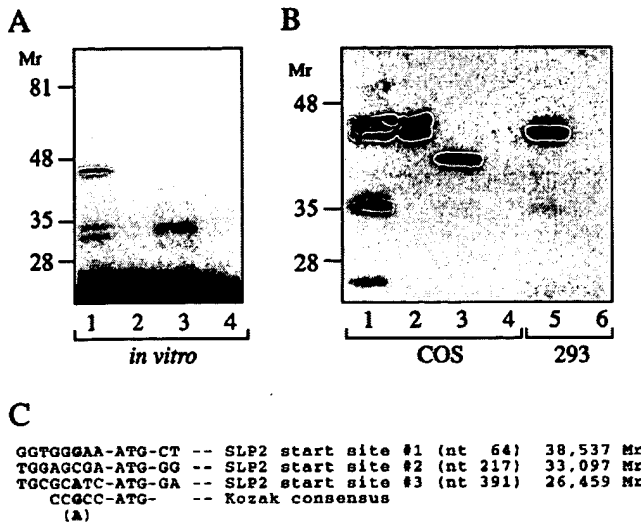


FIG. 7. Recombinant SLP-2 translated from the first ATG generates multiple bands identical to SLP-2 in COS cells. **A**, full-length SLP-2 cDNA in pCR2.1 was transcribed and expressed in the presence of 35 S-labeled methionine using a rabbit reticulocyte lysate. Shown is an autoradiograph of the expressed protein. Note the presence of a single protein product at $M_r \sim 45,500$ (calculated $M_r = 38,537$) when SLP-2 cDNA was used in the reaction (lane 1). When naked DNA was added in lanes 2 and 4 as controls without the vector, no transcription was evident. In lane 3, SLP-2 cDNA in pCR2.1 with a mutation that created a stop codon just downstream of the first ATG was used. **B**, constructs of full-length SLP-2 (initiator ATG at nt 64; calculated $M_r = 38,537$) or a truncated SLP-2 with its initiator ATG at nt 391 (calculated $M_r = 26,459$) were prepared and expressed in COS and 293T cells. All constructs placed an eight-residue FLAG epitope tag at the COOH terminus of the recombinant protein. Transiently transfected cells were then extracted with Triton X-100, and the soluble and pellet fractions were analyzed by Western blotting with anti-FLAG antibodies. The positions of the COS and 293T samples are as marked. Lanes 1 and 2, full-length FLAG-SLP-2 (soluble and pellet fractions, respectively). Lanes 3 and 4, truncated FLAG-SLP-2 (soluble and pellet fractions, respectively). Lanes 5 and 6, full-length FLAG-SLP-2 (soluble and pellet fractions, respectively) in 293T cells. Note that the full-length FLAG-SLP-2 generates a pattern identical to that in wild type COS and 293T cells, with the former showing bands at $M_r \sim 45,500$ and $44,600$ and occasionally at $M_r \sim 34,300$, while only the $M_r \sim 44,600$ band was present in the 293T cells. Approximately 60% of the FLAG-SLP-2 was inextractable in COS cells, while all was extractable in the 293T cells. By comparison, the truncated FLAG-SLP-2 construct appeared as a single band of $M_r \sim 40,500$. **C**, comparison of the three possible initiator sites in the SLP-2 gene, together with the predicted size of the resultant protein. All transcripts are expected to terminate at nt 1134.

In Fig. 6, just three SLP-2 immunoreactive bands were evident in erythrocytes. However, in the more heavily loaded gels shown in Fig. 8, many additional bands are apparent (α -j in Fig. 8A). Four of these bands (the most prominent) corresponded closely to those apparent in the transfected COS cells, at $M_r \sim 45,500$, $\sim 44,600$, $\sim 34,300$, and $\sim 26,300$. All other bands were larger than $M_r 45,500$. Under conditions of detergent extraction, only the major SLP-2 band at $44,600$ was soluble; all other bands (that collectively accounted for about 60% of the total SLP-2), including the major SLP-2 band at $45,500$, remained with the detergent-insoluble matrix (Fig. 8B). This insoluble fraction is operationally defined as the cytoskeletal matrix, although GPI-linked proteins may also remain insoluble under these conditions (31).

The presence of immunoreactive SLP-2 bands larger than $\sim 45,500$ suggested that some fraction of the protein might exist in covalent oligomeric complexes, or at least in complexes that are incompletely dissociated in SDS (such as occurs with glycoprotein dimers (32)). In preliminary experiments (data not shown), it was found that in freshly prepared human red cell ghosts two pools of SLP-2 appeared to exist: one strongly associated with the membrane that resisted salt-extraction and a

second pool that was fully extracted by 0.5 M KCl and that appeared to be composed of a high molecular weight complex involving still unidentified partners. In future studies, it will be important to identify the nature of the extractable oligomeric complex and its relationship to the more tightly associated membrane pool of SLP-2. To evaluate whether SLP-2 was also found in the cytosol of red cells, the hemolysate and membrane fractions from a fixed amount of red cells were analyzed by Western blotting; no SLP-2 was detected in the hemolysate under conditions in which the SLP-2 band was clearly discernible in the ghost (data not shown). In addition, the cellular distribution of spectrin, stomatin, and SLP-2 was examined in fresh human red cells by indirect immunofluorescent microscopy (Fig. 8C). In all cases, these proteins were present only beneath the plasma membrane, and in noncontacted cells they were approximately evenly distributed about the membrane. When two erythrocytes were in close contact, SLP-2 appeared to concentrate beneath the cell-cell contact sites, a property not shared by spectrin or stomatin.

SLP-2 Is a Peripheral Membrane Protein—While the structure of SLP-2 is consistent with its disposition as a peripheral membrane protein, its inability to be fully extracted by 0.5 M KCl suggested a tighter association with the bilayer than most skeletal proteins. To further explore this issue, the extractability of stomatin and SLP-2 with Triton X-100 or pH 11 NaOH was compared (Fig. 9). Triton extracted a portion of both stomatin and SLP-2. Conversely, stomatin was completely resistant to NaOH extraction (the hallmark of an integral membrane protein), while SLP-2 was completely extracted by such treatment. The NaOH extractability of SLP-2 confirms that it exists in red cells as a peripheral (but well attached) membrane protein.

DISCUSSION

The studies presented here identify SLP-2 as a novel member of the stomatin gene superfamily and reveal several unusual properties of this protein that may offer insights into its function. Distinguishing features include the following: 1) SLP-2 uniquely lacks a hydrophobic domain and functions as a peripheral (*versus* integral) membrane protein; 2) multiple SLP-2-related protein bands are evident on SDS-PAGE analyses of erythrocytes and other cells (most of which migrate more slowly than expected based on their calculated M_r); 3) SLP-2 is present in mature erythrocytes as well as in many if not all other types of tissues and cells; 4) SLP-2 partitions into a large oligomeric protein complex that is fully salt-extractable; 5) SLP-2 maps to the same chromosome as stomatin, although at a different locus (9p13 *versus* 9q34.1 for stomatin (33)); and 6) SLP-2 appears to concentrate in regions of erythrocyte membrane deformation or cell-cell contact. Collectively, these observations describe an unusual protein, reveal a novel and heretofore unrecognized component of the peripheral membrane skeleton of erythrocytes, and suggest novel and testable hypotheses as to its function.

The origin of the multiple sized SLP-2 protein bands evident in Western blots of erythrocytes, COS cells, and A431 cells remains uncertain. The observed bands fall into two categories: set 1, composed of approximately four bands of M_r $45,500$, $44,600$, $34,300$, and $26,000$, and set 2, a group of more variable bands above M_r $45,500$. Of the first group, the two largest bands (M_r $45,500$ and $34,300$) are the most abundant. In most cell types, just one of these two bands is present, although both appear in COS cells, A431 cells, and erythrocytes. Similarly, the smaller bands at $34,300$ and $26,000$ are absent in most tissues and cell types but are expressed (albeit in lesser amounts) in COS, A431, and red cells. The *in vitro* translation of SLP-2 cDNA also generates this same ensemble of protein

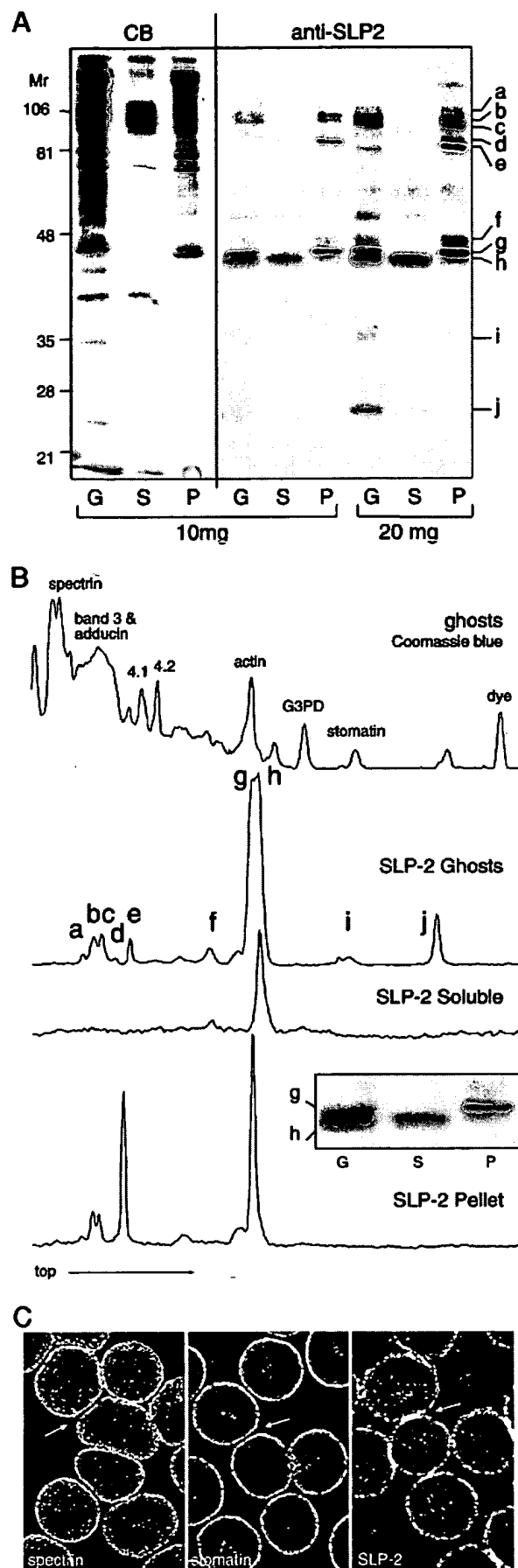


FIG. 8. SLP-2 associates with the Triton-insoluble cytoskeleton in erythrocyte ghosts. A, human erythrocyte ghosts (G) were extracted with Triton X-100 in isotonic buffer, and the presence of

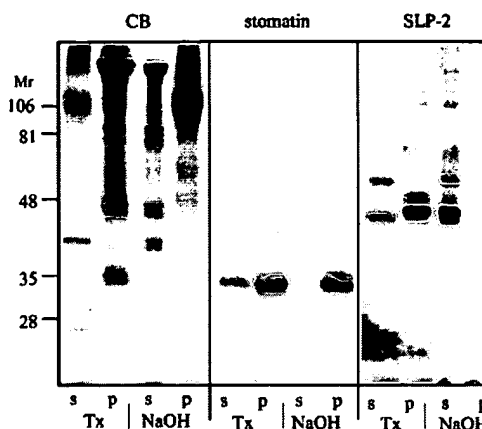


FIG. 9. SLP-2 is a peripheral membrane protein in red blood cells. To further explore the disposition of SLP-2 in the red cell membrane, its Triton extractability was compared with its ability to be extracted by NaOH at pH 11. Also compared in this assay was the extractability of stomatin under the same conditions. As before (Fig. 8), the majority of stomatin and SLP-2 was insoluble in Triton X-100 extracts. Conversely, when ghost membranes were extracted with pH 11 NaOH, a condition that removes all peripheral membrane proteins but does not extract the integral membrane proteins, stomatin was found to behave as an integral protein, while SLP-2 partitioned as a peripheral membrane protein. The panel on the left (CB) is Coomassie Blue-stained; the two panels on the right are Western blots with specific antibody to either human erythrocyte stomatin (stomatin) or human SLP-2. The positions of the molecular weight markers are as indicated.

bands as does FLAG-labeled recombinant SLP-2 when expressed in COS cells (but not 293T cells). While it remains possible that these smaller bands represent proteolytic products generated from the parent band at 45,500 M_r , we do not favor this interpretation. Instead, taking into consideration the consistency of these bands in three diverse cell types, their complete absence in other cell types, and their appearance after *in vitro* translation, we propose that these bands represent the products of tissue-specific alternative pre-mRNA splicing, alternative translation initiation at downstream AUGs, or both. Two candidate AUGs for such alternative initiation would be those at nt 217 and 391, the latter flanked by an excellent Kozak sequence. While future experimental work will be required to prove this conjecture, it is also worth noting that removal of the NH_2 -terminal portions of SLP-2, as would occur with initiation at either of the downstream AUGs, leads to a

SLP-2 in the soluble (S) and insoluble (P) fractions was examined by Western blotting. The Coomassie Blue-stained fractions (left panel) reveal the expected fractionation of proteins between the soluble and insoluble pools, with all of the cytoskeletal proteins remaining in the Triton-insoluble pool. Two prominent SLP-2 immunoreactive bands are present in ghosts (right panel, bands g and h; also the inset in B). The smaller band (M_r ~44,600) is Triton-soluble; the larger band (M_r ~45,500) is Triton-insoluble. At higher loadings (20 versus 10 μg), additional SLP-2 immunoreactive bands are observed at M_r values of 110,200 (a), 100,800 (b), 94,300 (c), 84,600 (d), 79,800 (e), 48,200 (f), 45,500 (g), 44,600 (h), 34,300 (i), and 26,300 (j). B, densitometric scans of the Coomassie Blue-stained ghosts and the SLP-2 Western blots of ghosts and the Triton-soluble and -insoluble fractions. Note that only band g (~44,500) is soluble; the rest remain with the Triton-insoluble pellet. Inset, enlarged view of the segregation of SLP-2 bands g and h between the soluble and insoluble Triton fractions. Over multiple determinations, $34 \pm 10\%$ (± 2 S.D.) of SLP-2 was Triton-soluble. By comparison, stomatin was $15 \pm 10\%$ (± 2 S.D.) Triton X-100-extractable in these experiments (Data not shown and Fig. 9). C, the intracellular distribution of SLP-2 in mature human erythrocytes was observed by indirect immunofluorescent microscopy (right) and was compared with the distribution of α IIb spectrin in these cells (left) and the distribution of stomatin (center). Note that substantially all of the detectable SLP-2 is arrayed with the membrane but in a more punctate pattern than is spectrin. It also appears to concentrate under points of membrane deformation or intercellular contact.

loss of detergent insolubility in COS cells (e.g. see Fig. 7) and thereby presumably altered intracellular function.

The nature of the higher molecular weight SLP-2 reactive bands is also enigmatic. Large complexes involving erythrocyte proteins have previously only been observed in cells that are oxidatively damaged (34, 35). Preliminary studies suggest that the large SLP-2-containing complexes that exist in fresh red cells might involve a covalent linkage (via a disulfide) of SLP-2 to another protein or proteins. The components of such a putative complex remain to be determined.

Given that stomatin's self-association appears to be mediated by the COOH-terminal portions of its sequence, a region sharing high homology to SLP-2 and predicted to be largely α -helical, an intriguing possibility that may speak to the role of SLP-2 is that SLP-2 forms mixed oligomers with stomatin. Stomatin is a much more abundant protein (~100,000 copies/cell) that exists by itself as large oligomers ($n = 9-12$) in the plasma membrane (20). The low stoichiometry of SLP-2 compared with stomatin requires that only a small subset of the total stomatin in the cell could be directly associated with SLP-2. At a measured ratio of about one molecule of SLP-2 for every 10-40 molecules of stomatin (in the red cell), we envision that each oligomeric stomatin complex might include one copy of SLP-2. Since stomatin oligomers might play a role in organizing cholesterol or sphingolipid-rich membrane rafts, along with the acylated and GPI-linked proteins typically associated with such rafts (36), a linkage to SLP-2 would provide a potential mechanism tying lipid rafts with their embedded proteins to the cytoskeleton.

If SLP-2 does interact with stomatin in red cells, several important implications follow. In both the dehydrated and overhydrated forms of hereditary stomatocytosis, there are defects in monovalent cation control and variable deficiencies in the level of stomatin. Recent data indicate that both dehydrated hereditary stomatocytosis and familial pseudohyperkalemia are linked to a gene at locus 16q23-qter (4), excluding disorders of stomatin or SLP-2 (both present on chromosome 9) as the cause of these disorders. However, the genesis of the overhydrated form of stomatocytosis remains unknown. This is the disorder with a most severe deficiency in stomatin, although stomatin mRNA is normal and appears to be made in normal amounts in afflicted individuals (3). While still unappreciated defects in the untranslated portions of the mRNA cannot be definitively ruled out, a more likely explanation of the available data suggests that in such patients, stomatin is made in normal amounts but is lost due to its rapid degradation, perhaps because it is incorrectly or inadequately assembled onto the membrane. If, as hypothesized, SLP-2 were required for its maintenance or attachment to the cytoskeleton, the loss of stomatin might follow defective SLP-2 function. It is also interesting to note in light of the apparent ease with which SLP-2 can be oxidatively cross-linked to itself or to other proteins, that the monovalent cation permeability of deformed red cells is also extremely sensitive to sulfhydryl cross-linking in the absence of reduced dithiothreitol (35). With respect to other disorders that might involve SLP-2, only speculation is possible at this point. Review of the human genome data base for other possible disorders that might map to 9p13 also proved unrevealing (37). Only about 3.4% of chromosome 9 has been sequenced at this point, and none of the disorders linked to this region of the genome offer any obvious relationship to the presumed functions of SLP-2. However, one interesting association is the close proximity of the genes for aquaporin 3 and aquaporin 7 to this region of chromosome 9. These genes are involved in water permeability regulation and cell volume control in a diverse array of tissues (38); an interesting conjecture

is that SLP-2 might be coordinately expressed with some of the aquaporin genes and play a role in their regulation. In future work, it will thus be important to examine the interaction of SLP-2 with stomatin or aquaporin and its potential role in organizing lipid rafts or monovalent cation permeability control.

Acknowledgments—The expert assistance of Paul Stabach and Amy Chang along with Drs. John Sinard, Deepti Pradhan, and Carol D. Cianci is gratefully acknowledged.

REFERENCES

- Stewart, G. W., Hepworth-Jones, B. E., Keen, J. N., Dash, B. C., Argent, A. C., and Casimir, C. M. (1992) *Blood* 79, 1593-1601
- Hiebl-Dirschmied, C. M., Entler, B., Glotzmann, C., Maurer-Fogy, I., Stratowa, C., and Prohaska, R. (1991) *Biochim. Biophys. Acta* 1090, 123-124
- Stewart, G. W., Argent, A. C., and Dash, B. C. (1993) *Biochim. Biophys. Acta* 1225, 15-25
- Delaunay, J., Stewart, G., and Iolascon, A. (1999) *Curr. Opin. Hematol.* 6, 110-114
- Stewart, G. W. (1993) in *Red Cell Membrane Antigens* (Tanner, M. J. A., and Anstee, D. J., eds) pp. 167-175, Balliere-Tyndall, London.
- Innes, D. S., Sinard, J. H., Gilligan, D. M., Snyder, L. M., Gallagher, P. G., and Morrow, J. S. (1999) *Am. J. Hematol.* 60, 72-74
- Zhu, Y., Paszty, C., Turetsky, T., Tsai, S., Kuypers, F. A., Lee, G., Cooper, P., Gallagher, P. G., Stevens, M. E., Rubin, E., Mohandas, N., and Mentzer, W. C. (1999) *Blood* 93, 2404-2410
- Snyers, L., Thines-Sempoux, D., and Prohaska, R. (1997) *Eur. J. Cell Biol.* 73, 281-285
- Huang, M., Gu, G., Ferguson, E. L., and Chalfie, M. (1995) *Nature* 378, 292-295
- Seidel, G., and Prohaska, R. (1998) *Gene (Amst.)* 225, 23-29
- Barnes, T. M., Jin, Y., Horvitz, H. R., Ruvkun, G., and Hekimi, S. (1996) *J. Neurochem.* 67, 46-57
- Rajaram, S., Sedensky, M. M., and Morgan, P. G. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 8761-8766
- You, Z., Gao, X., Ho, M. M., and Borthakur, D. (1998) *Microbiology* 144, 2619-2627
- Moore, R. B., and Shriver, S. K. (1997) *Biochem. Biophys. Res. Commun.* 232, 294-297
- Desneves, J., Berman, A., Dynon, K., La Greca, N., Foley, M., and Tilley, L. (1996) *Biochem. Biophys. Res. Commun.* 224, 108-114
- Ho, M. M., Nicolaou, A., Argent, A. C., and Stewart, G. W. (1997) *Biochem. Soc. Trans.* 25, 492S
- Snyers, L., Umlauf, E., and Prohaska, R. (1999) *FEBS Lett.* 449, 101-104
- Hiebl-Dirschmied, C. M., Adolf, G. R., and Prohaska, R. (1991) *Biochim. Biophys. Acta* 1065, 195-202
- Salzer, U., Ahorn, H., and Prohaska, R. (1993) *Biochim. Biophys. Acta* 1151, 149-152
- Snyers, L., Umlauf, E., and Prohaska, R. (1998) *J. Biol. Chem.* 273, 17221-17226
- Gallagher, P. G., and Forget, B. G. (1995) *J. Biol. Chem.* 270, 26358-26363
- Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
- Kodak International Biotechnologies, Ind. (1992) *Technical Bulletin* 1, Rochester, NY
- Devarajan, P., Stabach, P. R., Mann, A. S., Ardito, T., Kashgarian, M., and Morrow, J. S. (1996) *J. Cell Biol.* 133, 819-830
- Kennedy, S. P., Warren, S. L., Forget, B. G., and Morrow, J. S. (1991) *J. Cell Biol.* 115, 267-277
- Laemmli, U. K. (1970) *Nature* 227, 680-685
- Pierce (1994) *Pierce Catalog and Handbook*, p. T116, Pierce, Rockford, IL
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* 25, 3389-3402
- Kozak, M. (1984) *Nucleic Acids Res.* 12, 857-872
- Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J. (1994) *Nature* 368, 32-38
- Brown, D. A., and Rose, J. K. (1992) *Cell* 68, 533-544
- Bormann, B. J., Knowles, W. J., and Marchesi, V. T. (1989) *J. Biol. Chem.* 264, 4033-4037
- Westberg, J. A., Entler, B., Prohaska, R., and Schroder, J. P. (1993) *Cytogenet. Cell Genet.* 63, 241-243
- Schrier, S. L., and Mohandas, N. (1992) *Blood* 79, 1586-1592
- Hebbel, R. P., and Mohandas, N. (1991) *Biophys. J.* 60, 712-715
- Simons, K., and Ikonen, E. (1997) *Nature* 387, 569-572
- McKusick, V. A. (1997) *Online Mendelian Inheritance in Man (OMIMTM)*, Johns Hopkins University and the National Center for BioTechnology Information, National Library of Medicine, Baltimore and Bethesda, MD
- Agre, P., Brown, D., and Nielsen, S. (1995) *Curr. Opin. Cell Biol.* 7, 472-483
- Higgins, D. G., and Sharp, P. M. (1988) *Gene (Amst.)* 73, 237-244
- Chou, P. Y., and Fasman, G. D. (1978) *Adv. Enzymol. Relat. Areas Mol. Biol.* 47, 45-148
- Wang, Y., Chang, A., and Morrow, J. S. (1999) *Blood* 94, 189a

SeqServer[®]
biology in silico

ClustalW Results

Sequences

Help

Retrieval

BLAST2

FASTA

ClustalW

GCG Assembly

Phrap

Translation

Confidential -- Property of Incyte Corporation SeqServer Version 4.6 Jan 2002

☐ 789094CD1

☐ AF190167

CLUSTAL W (1.7) Multiple Sequence Alignments

Sequence format is Pearson

Sequence 1: 789094CD1 356 aa

Sequence 2: AF190167 356 aa

Start of Pairwise alignments

Aligning...

Sequences (1:2) Aligned. Score: 94

Start of Multiple Alignment

There are 1 groups

Aligning...

Group 1: Sequences: 2 Score:4464

Alignment Score 1975

CLUSTAL-Alignment file created [baajUaace.aln]

CLUSTAL W (1.7) multiple sequence alignment

789094CD1 MLARAARGHWGPFAEG--LSTGFWPRSGRASSGLPRNTVVLFPQQEAWVVERMGRFHRI
AF190167 MLARAARGTGALLLRGSLLASGRAPR--RASSGLPRNTVVLFPQQEAWVVERMGRFHRI

***** . : . * * : * ** *****

789094CD1 LEPGLNILIPVLDRIRYVQSLKEIVINVPEQSAVTLDNVTLQIDGVLYLRIMDPYKASYG
AF190167 LEPGLNILIPVLDRIRYVQSLKEIVINVPEQSAVTLDNVTLQIDGVLYLRIMDPYKASYG

789094CD1 VEDPEYAVTQLAQTTRSELGKLSXDKVFRERESLNASIVDAINQAADCWGIRCLRYEIK
AF190167 VEDPEYAVTQLAQTTRSELGKLSLXDKVFRERESLNASIVDAINQAADCWGIRCLRYEIK

789094CD1 DIHVPPRVKESMQMQVEAERRKRATVLESEGTRESAINVAEGKKQAQILASEAEKAEQIN
AF190167 DIHVPPRVKESMQMQVEAERRKRATVLESEGTRESAINVAEGKKQAQILASEAEKAEQIN

789094CD1 QAAGEASAVLAKAKAKAEAIRILAAALTQHNGDAAAASLTVAEQYVSFAFSKLAKDSNTILL
AF190167 QAAGEASAVLAKAKAKAEAIRILAAALTQHNGDAAAASLTVAEQYVSFAFSKLAKDSNTILL

789094CD1
AF190167

PSNPGDVTSMVAQAMGVYGALTKAPVPGTPDSLSSGSSRDVQGTASXDEELDRVKMS
PSNPGDVTSMVAQAMGVYGALTKAPVPGTPDSLSSGSSRDVQGTASLDEELDRVKMS

Submit sequences to:



A novel member of the STOMATIN/EPB72/mec-2 family, stomatin-like 2 (STOML2), is ubiquitously expressed and localizes to HSA chromosome 9p13.1

C.M. Owczarek,^a H.R. Treutlein,^b K.J. Portbury,^a L.M. Gulluyan,^a I. Kola^a and P.J. Hertzog^a

^aCentre for Functional Genomics and Human Disease, Monash Institute of Reproduction and Development, Monash University, Clayton, Victoria;

^bMolecular Modelling Group, Ludwig Institute for Cancer Research, P.O. Royal Melbourne Hospital, Parkville, Victoria (Australia)

Abstract. A cDNA encoding a novel second member of the Band7/stomatin-like/SPFH domain family in humans designated stomatin-like 2 (STOML2) has been isolated using the technique of cDNA Representational Difference Analysis. The STOML2 cDNA encoded a 356 amino acid residue polypeptide with a predicted molecular weight of 38.5 kDa. The predicted polypeptide sequence of STOML2 could be delineated into three major domains: an N-terminal α -helical region; a domain with significant similarity to a 172 amino acid region of the HSA stomatin polypeptide, composed of an alternating α -helical and β -sheet structure and a C-terminal domain that was mostly α -helical. The stomatin-like domain was observed in 51 other proteins with potentially diverse functions. Based on its homology to stomatin, STOML2 was predicted to be cytoplasmically located. However, unlike most of the other proteins containing stomatin-like domains, the predicted

STOML2 polypeptide does not contain a transmembrane region although the presence of N-myristoylation sites suggest that it has the potential to be membrane-associated. Northern blot analysis of a panel of poly(A)⁺ mRNA from normal human adult tissues showed that a single 1.3-kb mRNA transcript encoding STOML2 was ubiquitously expressed, with relatively higher levels in skeletal muscle and heart compared to other tissues. Comparison of the STOML2 cDNA sequence with human genomic DNA indicated that the gene encoding STOML2 was 3,250 bp long and consisted of ten exons interrupted by nine introns. We have mapped STOML2 to HSA chromosome 9p13.1, a region that is rearranged in some cancers and thought to contain the gene responsible for acromesomelic dysplasia.

Copyright © 2001 S. Karger AG, Basel

Stomatin or erythrocyte band 7 protein (EPB72) is a 31-kDa integral membrane palmitoylated phosphoprotein with a monotopic membrane protein structure (Hiebl-Dirschmied et al., 1991; Stewart et al., 1992). The normally ubiquitously ex-

pressed stomatin was thought to play a role as a regulator of ion channels because the protein is absent or partially deficient in the erythrocytes of patients with overhydrated hereditary stomatocytosis (Stewart et al., 1993; Stewart 1997). However, its physiological role still remains undefined as the unaltered mRNA encoding the stomatin polypeptide is present in these patients (Stewart et al., 1993). Furthermore, mice with a homozygous null mutation in the murine orthologue of the stomatin gene are phenotypically normal and have normal red blood cells (Zhu et al., 1999). Recent studies suggest that stomatin may have a scaffolding function based on its topological similarity to caveolin (Snyers et al., 1998). In humans, a cDNA clone encoding a brain-specific stomatin homologue, SLP1 (Seidel and Prohaska 1998) has been identified. In *C. elegans*, genes encoding several stomatin-like proteins have been identified (mec-2, unc-1, unc-24) and genetic mutants have defects in

Supported by the National Health and Medical Research Council of Australia. The Monash Institute of Reproduction and Development was established and is supported by Monash University.

Received 17 July 2000; revision accepted 12 October 2000.

Request reprints from Dr. Catherine Owczarek, Centre for Functional Genomics and Human Disease, Monash Institute of Reproduction and Development, Monash University, Clayton, Victoria 3168 (Australia); telephone: 61-3-9594 7223; fax: 61-3-9594 7211; e-mail: catherine.owczarek@med.monash.edu.au

Present address of L.M.G.: Howard Florey Institute of Medical Research, Parkville, Victoria 3050 (Australia)

either mechanosensation or locomotion (Huang et al., 1995; Barnes et al., 1996; Rajaram et al., 1998).

In this report we describe the cloning and chromosomal localisation of a novel member of the stomatin-like domain family in humans, stomatin-like 2 (STOML2). We also demonstrate the expression of this novel gene in various normal adult human tissues and examine the relationship of its predicted protein product with other members of the stomatin-like domain family.

Materials and methods

Cell culture

The parental CHO-K1 cell line was obtained from the American Type Culture Collection (ATCC). The Hamster-HSA chromosome 21-containing hybrid cell lines 21q+ and 72532x6 were obtained from D. Patterson (Eleanor Roosevelt Institute, Denver, CO). The HSA chromosomal complement of the hybrids has been described elsewhere (Graw et al., 1995). Cell line 72532x6 was grown in RPMI 1640 medium supplemented with 5% dialysed fetal calf serum, 100 units/ml penicillin, 100 µg/ml streptomycin. CHO-K1 and 21q+ cultures were grown under the same conditions except the medium was supplemented with 2.3 mg/ml proline.

cDNA Cloning and sequence analysis

Cells were grown to mid-log phase, harvested and poly(A)⁺ mRNA prepared as previously described (Gonda et al., 1982). Preparation of double-stranded cDNA was carried out according to the method described in Braun et al. (1995).

In order to isolate genes located on HSA chromosome 21q22.2 → q22.3, cDNA RDA was performed using tester cDNA derived from a hamster-HSA somatic cell hybrid containing an intact HSA chromosome 21 (72532x6) and driver cDNA from a CHO-human hybrid cell line (21q+) that was identical except that it lacked the region of HSA chromosome 21 from q22.2 → q22.3. The cell lines were grown at the same time and to a similar level of confluency, and the cDNA produced from each cell line was of similar size range and intensity. RDA subtractive hybridization and PCR was accomplished essentially as described (Hubank and Schatz, 1994). PCR products were gel-purified, digested with *DpnII* and ligated into pBluescript KS II+ (Stratagene) digested with *BamHI* and dephosphorylated with calf intestinal alkaline phosphatase (Roche).

Automated sequencing was carried out using plasmid templates using T3 and T7 primers on a Prism Reaction Ready Dye Termination mix on an ABI automated sequencer at the Monash University Microbiology Sequencing Facility. Similarity searches were carried out using the BLAST programs at NCBI.

Library screening

In order to generate a probe for library screening a contiguous nucleotide sequence of 1,200 bp encoding STOML2 was generated from several Expressed Sequence Tags (ESTs) extracted from the Genome Database (GDB) and assembled using the SequencherTM Sequence Analysis version 3.0 package (Gene Codes Corporation). A 1,200-bp PCR product was then generated using Taq polymerase (Promega) using human lung cDNA as a template, subcloned into pGEM-T (Promega) and its sequence confirmed. 1 × 10⁶ clones from a λgt11 human fetal brain cDNA library (Clontech) were screened with the cloned STOML2 PCR product using standard techniques. After three rounds of screening twenty clones were isolated. Preparations of λ DNA were carried out using standard methods. The DNA preparations were digested with *EcoRI*, the inserts excised and ligated into pBluescript KS II+ digested with *EcoRI* and dephosphorylated with calf intestinal alkaline phosphatase (Roche).

Northern blot analysis

Northern blot analysis of poly(A)⁺ mRNA (3 µg) from cell lines was carried out as previously described (Owczarek et al., 1997). The membranes were hybridized with a variety of ³²P-labeled RDA cDNA clones, then finally with a ³²P-labeled 1.1-kb glyceraldehyde phosphate dehydrogenase (GAPDH)

cDNA probe as described previously (Owczarek et al., 1997). Northern blots containing human tissues were obtained from Clontech and hybridized as described above. The amounts of poly(A)⁺ mRNA in each lane was confirmed by hybridisation using a ³²P-labeled β-actin probe. After hybridisation and subsequent washing in 0.1 × SSC, 0.1% SDS at 65°C, signals were visualised by autoradiography onto Kodak BioMax film.

Southern blot analysis

10 µg of genomic DNA was digested with *BamHI* (Promega) according to the manufacturer's instructions, electrophoresed in a 0.8% agarose gel and then transferred to GeneScreen Plus membrane (Dupont) in 0.4 M NaOH.

Human chromosome mapping

Chromosome assignment was obtained by Southern analysis of a *BamHI* digested monochromosomal human/rodent somatic cell hybrid mapping panel (Coriell Institute of Medical Research, Camden, NJ). Specific signals in DNA from somatic cell hybrids that were of identical size to those found in control human genomic DNA, but not in mouse or hamster DNA, indicated the gene locus at a single chromosome resolution. The Genebridge 4 panel of radiation hybrids was obtained from the UK HGMP Resource Centre. PCR was performed on 50 ng of DNA using gene-specific primers and buffers from the PCR Optimizer Kit (Stratagene) under conditions experimentally determined for each primer pair. Specific PCR products were obtained in human control DNA but not in hamster or mouse DNA. Linkage analysis was carried out using the RhyME suite of programs (<http://hgmp.mrc.ac.uk/>) that uses RADMAP to predict the location of the marker in relation to the human 1998 International Gene Map. Only LOD scores greater than 3 were taken into account. The results were compared with previously defined markers to obtain the relative cytogenetic positions.

Secondary structure prediction and sequence alignment

The secondary structure of the STOML2 polypeptide was predicted using the JPRED program (Cuff et al., 1998). JPRED is a method for protein secondary structure prediction, which is based on the consensus of a number of current secondary structure prediction algorithms. The predictions obtained from JPRED were also compared with an independent method PSIPRED (Jones, 1999). It was found that both methods resulted in very similar predictions. In the remainder of the paper, only results from the JPRED prediction method will be presented and discussed. JPRED was also used to align amino acid sequences via an interface to CLUSTALW (Thompson, 1994).

Prediction of helical transmembrane domains

The helical transmembrane domains and topology of STOML2, mec-2 and stomatin were predicted using the MEMSAT2 method (Jones, 1998), which is based on the MEMSAT prediction algorithm (Jones et al., 1994).

Protein domain analysis

Potential protein domains were predicted using the ProDom database and server (Corpet et al., 1999) version 99.2 from <http://protein.toulouse.inra.fr/prodom.html>. In our analysis we only took those domain predictions into account that showed more than one homologue in the ProDom database. The ProDom analysis is also useful for finding homologues of the various domains that sometimes can be linked to a certain function of the protein.

Results

Generation and characterisation of cDNA clones obtained by Representational Difference Analysis

The technique of cDNA Representational Difference Analysis (cDNA RDA) (Hubank et al., 1994) is a subtractive PCR-based procedure that has been used to isolate genes that are differentially expressed between two mRNA populations. This technique was used in an effort to isolate a novel type I IFN signaling molecule (Hertzog et al., 1994; Raz et al., 1995; Holland et al., 1997) located in region q22.2 → q22.3 of HSA chro-

mosome 21. cDNA RDA was performed using tester cDNA derived from a hamster-human somatic cell hybrid containing a complete HAS chromosome 21 (72532x6) and driver cDNA from a hamster-human hybrid cell hybrid that was identical but had a deletion of HSA chromosome 21 from q22.2→q22.3 (21q+). Three rounds of RDA subtractive hybridization and PCR were performed. Partial cDNA clones corresponding to 14 different genes resulted from the final round of selection. Five partial cDNA clones encoding HSA chromosome 21-specific genes mapping distal and one mapping proximal to the 21q+ breakpoint were isolated: COL6A2, COL6A1, SMT3H1, CBS, inward rectifier K⁺ channel; and TIAM1 respectively. Northern blot analyses (data not shown) showed that these transcripts were expressed only in the 72532x6 cell line. Three differentially expressed genes of hamster origin were identified (cytochrome c-oxidase, Hox, ribosomal protein L9). A 2.1-kb cDNA clone was isolated that was not present in the 72532x6 cell line, human or hamster genomic DNA. A PCR product was detected in *E. coli* DNA using specific primers indicating that this clone was most likely of bacterial and not mammalian origin. Interestingly, this sequence was identical to putative HSA cDNAs encoding cerebrin-50 and the identical MYT2 (accession numbers 576853 and NM 003871 respectively), and highly homologous to origin of replication genes from several bacterial species.

Southern blot analyses of the Coriell panel of human-rodent monochromosomal hybrids using the four remaining cDNA clones showed that they did not map to HSA chromosome 21. Instead they mapped to HSA chromosome 9p13.2 (TPM2) or HSA Y chromosome (RPS4Y). Two cDNAs (4-7 and 4-8) that did not correspond to any known genes were also isolated. Clone 4-7 localised to HSA chromosome 9 by Southern blot analysis of the Coriell panel and to 9p13.1 (closest marker AFM136xc5, LOD 4.173) using the MRC HGMP Resource panel of radiation hybrids (Gyapay et al., 1996; Hayes et al., 1996). Clone 4-7 was identical to several HSA ESTs (GDB Accessions numbers ALO40260, AA429498, W42808, AA411191) but as they all contained an Alu repetitive sequence further analysis of clone 4-7 was discontinued.

Southern blot analysis of clone 4-8 indicated that a specific hybridizing band in 72532x6 genomic DNA was also present in the human control lane confirming that it was of human origin (data not shown). An approximately 1.3-kb mRNA transcript encoding this cDNA clone was abundantly expressed in the tester cell line 72532x6 (data not shown). Clone 4-8 was localised to HSA chromosome 9 using Southern blot analysis of the Coriell DNA panel. BLASTN analysis of the GDB database indicated that clone 4-8 was identical to several human and mouse ESTs and to regions in a HSA chromosome 9 P1 clone 11659 (GDB Accession number AC004472) but not to any known genes. BLASTX analysis found that the partial predicted amino acid sequence of clone 4-8 was similar to regions in a large number of proteins including HSA (and MMU) erythrocyte band 7/stomatin (EPB72), HSA stomatin like protein 1 (SLP1), *C. elegans* proteins mec-2, unc-1, unc-24 and putative protein f30a10.5 (Accession number Z81072), and stomatin-like proteins from various eubacteria and archaeobacteria. The partial predicted amino acid sequence was also distantly relat-

```

1  TCCTGGTTCGGAGGTCGCTGCGGGGTGGGAAATCTGGCCGCGCGCGCGGGGCAC
   M L A R A A R G T
61  TGGGGCCCTTTTGTCTGAGGGGCTCTCTACTGGCTTCTGGCCGCGCTCCGCGCGCGCTC
   G A L L L R G S L L A S G R A P R R A S
121 CTCTGGATTGCCCGGAAACACCGTGGTACTGTTCTGCTGCCGAGCAGGAGGCGCTGGTGGT
   S G L P R N T V V L F V P Q Q R A W V V
181 GGAGCGAATGGGCGGATTCACCGGATCTCGGAGCTGGTTTGAACATCTCTCATCCCTGT
   E R M G R F H R I L E P G L N I L I P V
241 GTTAGACCGGATCCGATATGTGTCAGAGTCTCAAGAAATGTTCATCAAGTGCCTGAGCA
   L D R I R Y V Q S L K E I V I N V P E Q
301 GTCCGCTGTGACTCTCGACAATGTAACTCTGCAAAATCGATGGAGTCTTTACCTGCGCAT
   S A V T L D N V T L Q I D G V L Y L R I
361 CATGGACCTTTACAAGGCAAGCTACGGTGGAGGACCTGAGTATCGCTCACCACGCT
   M D P Y K A S Y G V D P E Y A D T Q L
421 AGCTCAACAACCATGAGATCAGAGCTCGGCAAACTCTCTGGACAAAGTCTTCCGGGA
   A Q T T M R S E L G K L S L D K V F R E
481 ACGGAGTCCCTGAATGCCAGCATTTGGATGCCATCAACCAAGCTGCTGCTCGGGG
   R E S L N A S I V D A I N Q A A D C W G
541 TATCCGCTCCCTCCGTTATGAGATCAAGGATATCCATGTGCCACCCGGGTGAAAGATC
   I R C L R Y E I K D I H V P P R V K E S
601 TATGAGATGAGGTCAGGTCAGGCGCGGAAACAGGCCACAGTCTTAGAGTCTGAGGG
   M Q M Q V E A E R R K R A T V L E S E G
661 GACCCGAGATCGGCCATCAATGTGGCAGAGGGAAGAACAGGCCACAGTCTTGGCTC
   T R E S A I N V A E G K Q A Q I L A S
721 CGAAGCAGAAAGGCTGAACAGATAAATCAGGAGCAGGAGGAGGAGGAGTCTGCTGGC
   E A E K A E Q I N Q A A G E A S A V L A
781 GAAGGCCAAGGCTAAAGCTGAAGCTATTTCGAATCCTGGCTGTCAGCTCTGACACACATAA
   K A K A K A E A I R I L A A A L T Q H N
841 TGGAGATGACGAGCTTCACTGACTGTGGCCGAGCAGTATGTACGCGCTTCTCCAACT
   G D A A A S L T V A E Q Y V S A F S K L
901 GGCCAAGGACTCCAACTACTCTACTGCCCTCCAACTCGGAGTGTACACGACATGGT
   A K D S N T I L L P S N P G D V T S M V
961 GGCTCAGGCCATGGGTGTATATGGAGCCCTACCAAGGCCCAAGTGTCCAGGAGTCCAGA
   A Q A M G V Y G A L T K A P V P G T P D
1021 CTCACCTCCAGTGGGAGCAGCAGAGATGTCCAGGGTACAGTCTGATGAGGA
   S L S G S S R D V T C G T D A S L D E E
1081 ACTTGATCGAGTCAAGATGAGTATTAGTGGAGCTGGGCTTGGCCAGGAGTCTGGGACAAAG
   L D R V K M S
1141 GAAGCAGATTTCTCTGATCTGCTGCTAGCTTCCCTGCCAAGATTTTGGTTTATTTT
1201 TTATTGTAACCTTTAGTGTGTATAAATACTCACCAGTGGCAAAAAAAAAAAAAAAAAAAAA

```

Fig. 1. Nucleotide sequence of STOML2 cDNA. The positions of the exon-intron boundaries are indicated with arrows. The polyadenylation signal is underlined. The sequence has been deposited under GDB accession number AF282596.

ed to several proteins including prohibitin, flotillin, and the Hflc/Hflk complex in *E. coli*. The cDNA encoded by RDA clone 4-8 therefore represented a novel gene and has been designated stomatin-like 2 (STOML2).

Sequence of STOML2 cDNA

In order to isolate a full-length cDNA clone encoding STOML2, a λ gt11 human fetal brain cDNA library (Clontech) was screened. The largest clone (STOML2-6.5) had an insert of 1,260 bp. This was consistent with an expected STOML2 mRNA transcript size of approximately 1.3 kb. The cDNA sequence of STOML2 is shown in Fig. 1. The initiator methionine has been designated at nucleotide position 35 and the first in-frame stop codon at position 1105. A stop codon immediately upstream of the proposed initiator methionine and a Kozak consensus sequence were not present (Kozak, 1987) and hence we cannot rule out a translational start site further upstream. The proposed open reading frame of 1,068 nt encoded a 356

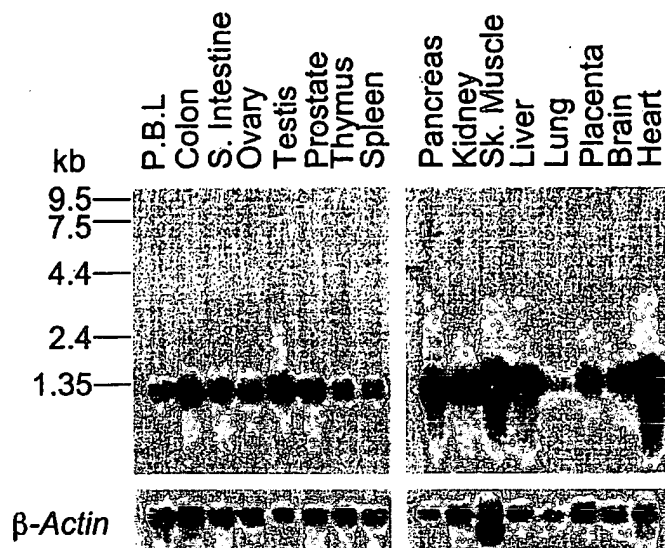


Fig. 2. Northern blot analysis of STOML2 mRNA expression in poly(A)⁺ RNA from a range of normal human adult tissues.

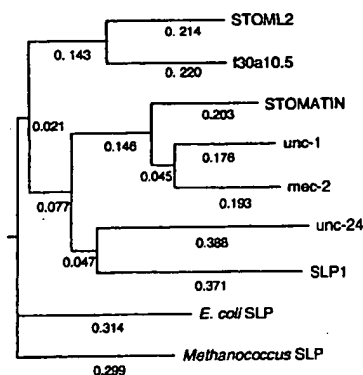
amino acid polypeptide with a predicted molecular mass of 38.5 kDa. A polyadenylation sequence was present at nucleotides 1222 to 1227. The sequence data has been deposited under GDB accession number AF282596.

Northern analysis of STOML2 mRNA expression

When cDNA clone 4-8 was hybridized to a Northern blot containing poly(A)⁺ mRNA from a range of normal human adult tissues a single mRNA transcript of approximately 1.3 kb was detected in all tissues (Fig. 2). Relatively higher levels of STOML2 mRNA were detected in skeletal muscle and heart compared to other tissues. When this filter was subsequently rehybridized with a full-length STOML2 cDNA probe an identical pattern of transcripts was observed.

Chromosomal localisation of the STOML2 gene

To map STOML2 more precisely we used the Genebridge 4 panel of radiation hybrids. The RhyME radiation hybrid mapping program predicted that STOML2 was most likely to be located between markers AFMa044ta5 and AFMa1yd1. The 2-point analysis indicated that the markers with the highest LOD scores were AFM136xc5 (LOD 6.216) and AFMa044ta5 (LOD 3.967). These markers corresponded to the cytogenetic position 9p13.1. This assignment was identical to clone 4-7 indicating that they probably arose from the same fragment of chromosome 9 present in the 72532x6 cell line. The XRCC9 gene (current symbol FANCG), which is present on the same P1 clone as the STOML2 gene, has been assigned to 9p13 by FISH (Liu et al., 1997). We also tested the chromosomal location of the XRCC9 gene using primers designed to amplify a PCR product from nucleotides 2295–2495 in the published XRCC9 mRNA sequence (Accession number NM_004629). The results were identical to those obtained for the STOML2 gene.



STOML2	1	MLARAARGTGALLRG--SLLASGRAPRRASSG----	LPRMTVVLFVQ	43
f30a01.5	1	MALTNRLLMNSSALLRGAILFRSSTLPVAVTSSRQAAHNTVINFPVQ		50

STOML2	44	QEAUVVERMGRFHRILEPGLNLIPLVDRIYVQSLKEIVINVEQSAVT		93
f30a01.5	51	QEAUVVERMGRFHRILEPGLNLIPLVDRIYVQSLKEIVINVEQSAVT		100

STOML2	94	LDNVTLQIDGVLYLRIMDPYK---ASYGVDEPEYAVTQLAQTTMRSELG		140
f30a01.5	101	IDNVQLRLDGVLYLRVDFPYKACDASYGVDDPEFAVTQLAQTTMRSEVGK		150

STOML2	141	LSLDKVFRRERESLNASIVDAINQADCGWIRCLRYEIKDIHVPPRVKESM		190
f30a01.5	151	INLDTVFKERELNENIVFAINKASAPWGIQCMRYEIRDMQMPSKIQEAM		200

STOML2	191	QMQVEAERRKRATVLESEGTRESAINVAEGKKQAQILASEAEKAEQINQA		240
f30a01.5	201	QMQVEAERRKRATVLESEGTRESAINVAEGKKQAQILASEAEKAEQINQA		250

STOML2	241	AGEASAVLAKAKAKAEAIRILAAALQTHNGDAAASLTVAEQYVSAPSKLA		290
f30a01.5	251	KGEAEAVILKAESRAKAIERIALALEKDGQANAAGLTVAEQYVGAQNL		300

STOML2	291	KDSNTILLPSNPGDVTSMVAQAMGVYGAITKAPVPGTDFSLSSGSSRDVQ		340
f30a01.5	301	KESNTVVLPAVLSDPGSMVQALAVY-----DSLS-----		330

STOML2	341	GTDAASLDEELDRVKMS	356	
f30a01.5	331	-----NKKK	334	

Fig. 3. (A) ClustalW analysis of a range of proteins related to STOML2 displayed as a phenogram. The distance between the nodes is shown in phylogenetic units. A value of 0.1 corresponds to approximately a difference of 10% between two sequences. (B) Pairwise alignment of the amino acid sequences of STOML2 and f30a10.5. Asterisks (*) indicate amino acid identity and dots (.) indicate amino acid similarity.

Secondary structure predictions of the deduced STOML2 polypeptide and comparison with the deduced polypeptide sequence of HSA stomatin and C. elegans mec-2

ClustalW analysis (Fig. 3A) of several proteins shown to be related to STOML2 by BLASTX analysis indicated that the predicted STOML2 polypeptide was most similar to a putative polypeptide encoded by the *C. elegans* gene f30a10.5 (Accession number Z81072). Figure 3B shows an alignment of the amino acid sequences of STOML2 and f30a10.5. Despite the high level of amino acid identity and similarity, 55% and 15% respectively, between these two proteins we carried out further detailed structural comparisons between STOML2, mec-2 and stomatin. This was because these proteins and their cognate genes are relatively well characterized whilst the f30a10.5 gene was predicted from genomic sequence using GeneFinder. Figure 4 shows an alignment of the amino acid sequences of the *C. elegans* stomatin-like protein mec-2 (Huang et al., 1995),

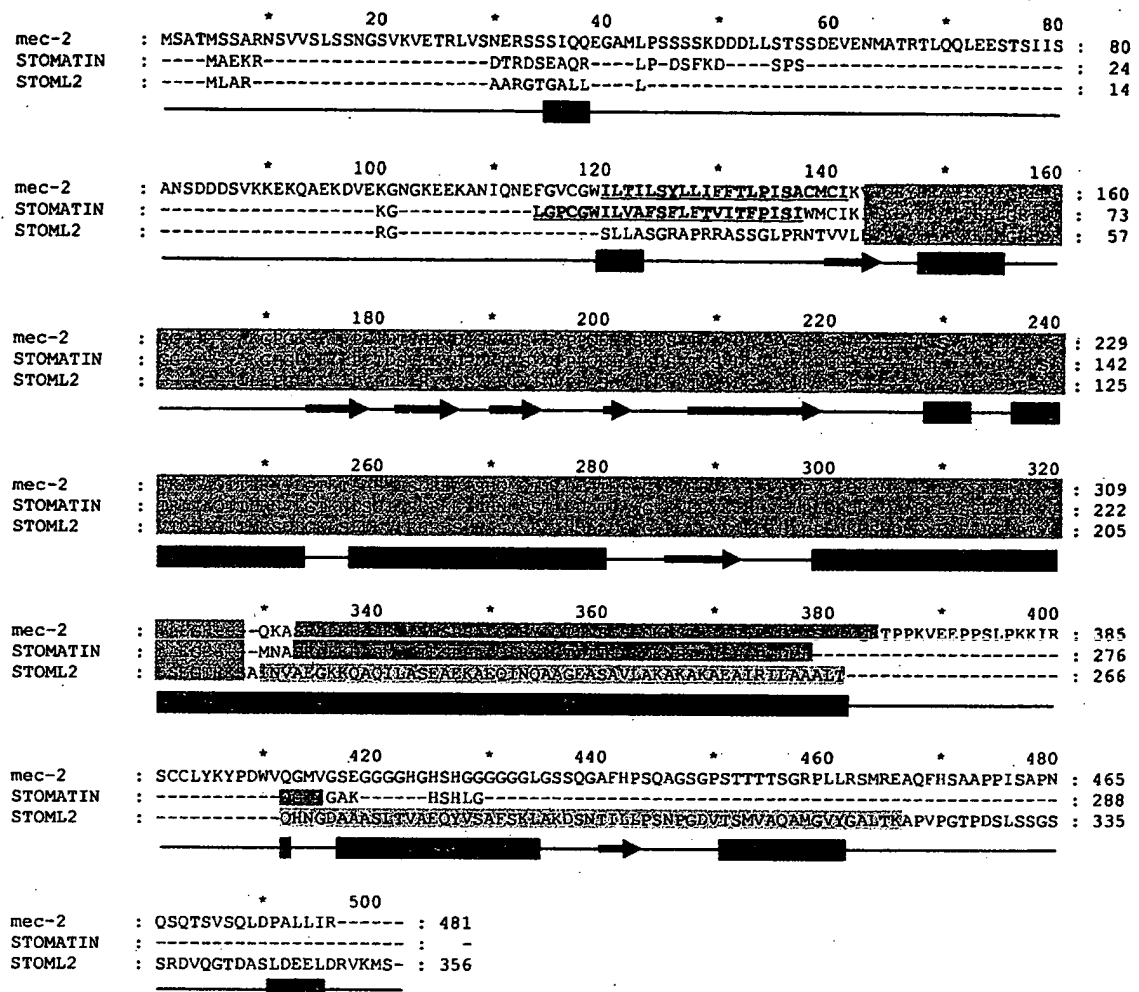


Fig. 4. Predicted secondary structure and domain organisation of STOML2 and sequence alignment with mec-2 and stomatin. Amino acids are colour-coded according to their hydrophobicity with red, blue and black indicating hydrophobic, hydrophilic, and intermediate residues, respectively. The secondary structure prediction of the STOML2 polypeptide is indicated schematically below the STOML2 amino acid sequence. Orange boxes and blue arrows show the position of the predicted helices and strands, respectively. Amino acids belonging to the predicted transmembrane domains are printed in bold and underlined. Sequence areas which belong to predicted ProDom domains are highlighted by coloured backgrounds: grey for domain PD001059, light yellow for domain PD022878, and green for domain PD005343.

HSA stomatin and STOML2 as well as the result of a secondary structure prediction for STOML2 and the prediction of helical transmembrane domains for all three sequences. As can be seen in Fig. 4 approximately two-thirds of the sequence of STOML2 align well with the other two sequences. Only the N- and C-terminal areas differ significantly. STOML2 does not have a transmembrane domain, whereas the other two proteins are clearly partitioned into N-terminal, transmembrane and intracellular C-terminal domains. The N-terminal domain of stomatin has been experimentally shown to be intracellular resulting in a hairpin-like arrangement of the protein in the plasma membrane (Salzer et al., 1993). The N-terminal domains of stomatin and mec-2 are completely different in length and amino acid composition. It can therefore be assumed that they are involved in quite different functions. In contrast, the intracellular

lar C-terminal domains of all three proteins consist of at least three distinct domains: (i) a highly conserved domain (residues 40–213 in STOML2), which was also identified in our protein domain search using the ProDom server as domain PD001059, (ii) a consecutive domain (215–321 in STOML2) which was identified as ProDom domain PD022878 and (iii) an unidentified short domain at the end of the C-terminal end (322–356 in STOML2). Domain PD001059 has been observed in 51 proteins in the database and occurs mostly in the Band 7 family of proteins, HflC, HflK and prohibitin. According to our secondary structure predictions this domain would most likely fold as a mixed α and β type fold. Domain PD022878 has been observed in four proteins in the ProDom database, none of which so far have a known function. Mec-2 and stomatin do not have a domain PD022878, but instead show a domain identi-

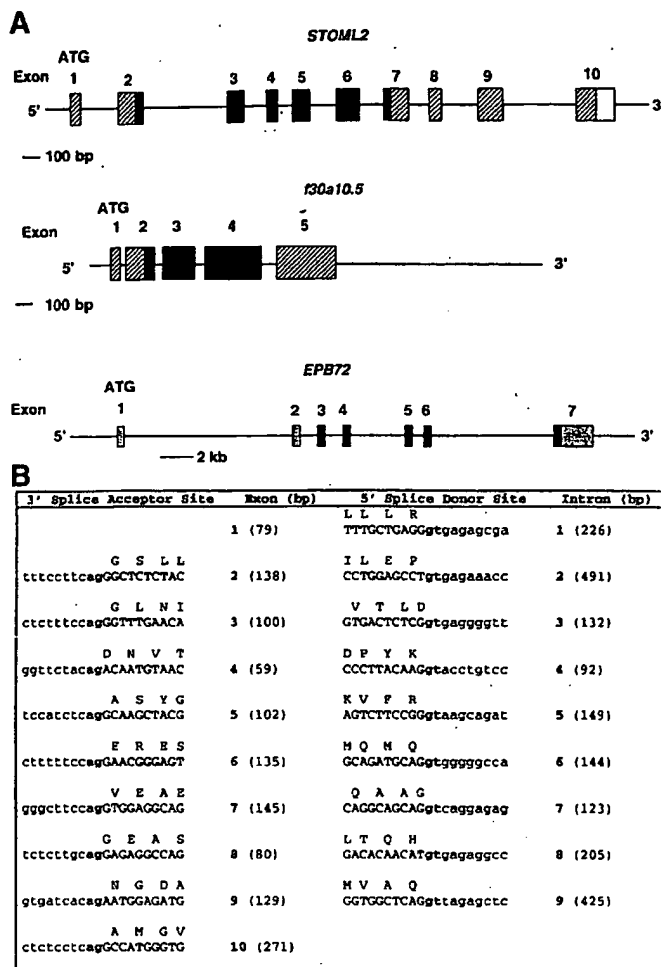


Fig. 5. (A) Intron-exon structure of the STOML2 gene, the f30a10.5 gene and the EPB72 gene. Exons are shown as boxes with the exon number indicated. Exons encoding regions of high amino acid homology between the STOML2, putative f30a10.5 and STOMATIN polypeptides are black. Exons coding for similar amino acids between the STOML2 and putative f30a10.5 polypeptide are shaded with diagonal lines. **(B)** Intron-exon boundaries of the STOML2 gene.

fied as PD005343 which is observed 15 times in the database and seems to be more specific to members of the Band 7 family.

The deduced STOML2 protein sequence was searched for protein motifs using the PROSITE Database of protein families and domains. The following motifs were found: two potential cAMP/GMP-dependent protein kinase phosphorylation sites at amino acid positions 26–29 and 200–203; four potential protein kinase C phosphorylation sites at amino acid positions 21–23, 78–80, 133–135 and 335–337; seven potential casein kinase II phosphorylation sites at amino acid positions 78–81, 156–159, 203–206, 229–232, 77–280, 335–338, 345–348; six potential N-myristoylation sites at positions 16–21, 31–36, 209–214, 14–319, 326–331, 341–346 and two potential N-glycosylation sites at positions 96–99 and 154–157.

Intron-exon organization of the STOML2 gene and comparison with the f30a10.5 and stomatin (EPB72) genes

The intron-exon boundaries of the STOML2 gene were determined by comparing the genomic sequence contained in the HSA chromosome 9 P1 clone 11659 with the STOML2 cDNA sequence. The STOML2 gene was 3,250 base pairs long and consisted of 10 exons interrupted by 9 introns (Fig. 5A). Comparisons of the exon-intron boundaries with the consensus splice sequences (Shapiro and Senepathy, 1987) are given in Fig. 5B. Interruption of the coding sequence by introns occurred between codons (class 0) and also after the first nucleotide (class I) (exons 3 and 7) (Sharp, 1981). All the splice boundaries were consistent with the GT...AG rule (Breathnach and Chambon, 1981). The exons ranged in size from 59 to 270 bp.

Since both the STOML2 and stomatin predicted polypeptides contained regions of significant amino acid similarity and the EPB72 gene is also located on HSA chromosome 9 (9q) it was of interest to compare the structure of their genes to see if a conserved arrangement reflected a common evolutionary origin. Conservation of intron-exon boundaries has been reported for the genes encoding SLP1, *C. elegans* unc-24 and EPB72 (Seidel et al., 1998). The intron-exon structure of the putative *C. elegans* orthologue encoded by the predicted gene f30a10.5 was also compared. When the f30a10.5 and STOML2 genes were compared only exons 1 and 2 were similar in their size and in the position of the splice junctions. The intron-exon structure of the STOML2 gene did not appear to be highly conserved with the stomatin gene (EPB72) with respect to gene size, intron-exon boundaries, phasing or relative position of conserved amino acid residues within a given exon even in regions of amino acid identity (Fig. 5A).

Discussion

The technique of cDNA Representational Difference Analysis (Hubank et al., 1994) is a subtractive PCR-based procedure used to isolate genes that are differentially expressed between two mRNA populations (Braun et al., 1995; Gress et al., 1997; Lewis et al., 1997; Wada et al., 1997). We used this technique to identify genes specifically expressed in the region q22.2→q22.3 of HSA chromosome 21 as we were interested in isolating a type I IFN signalling molecule that resides in this region (Hertzog et al., 1994; Raz et al., 1995; Holland et al., 1997). Genes localised on HSA chromosome 21q22.2→q22.3 were isolated validating this technique for cloning genes at particular chromosome positions. A total of 14 different cDNAs were isolated which is similar to the number of genes isolated by several other RDA experiments reported in the literature (Hubank et al., 1994; Braun et al., 1995; Niwa et al., 1997). Of the 14 cDNAs isolated nine were of human origin. However, of these nine, five were located on either 21q22.2 or 21q22.3, and one proximal to the 21q+ breakpoint. Three cDNA clones localised to HSA chromosome 9 and one to the HSA Y chromosome. These non-HSA 21 cDNAs were abundantly expressed and were selected during the RDA screen from contaminating fragments of the HSA 9 and Y chromosomes in the 72532x6 tester cell line.

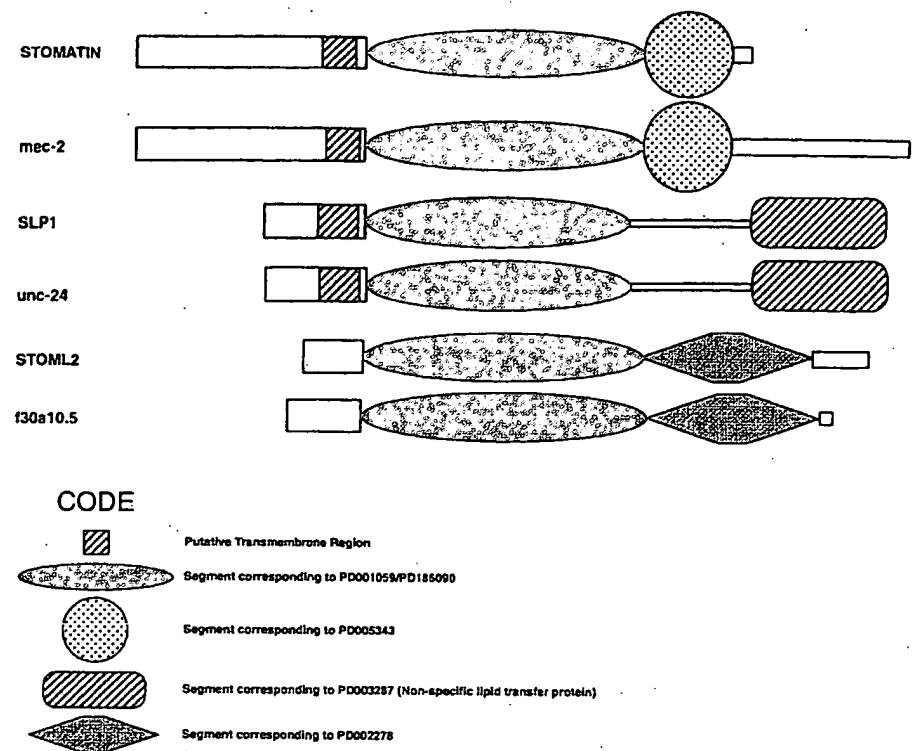


Fig. 6. Arrangement of domains in proteins related to STOML2. The stomatin-like/SPFH domain corresponds to ProDom domains PD001059/PD186090

Of the three clones localised on HSA chromosome 9 only one, clone 4-8, was novel. A full-length cDNA was subsequently isolated and found to contain a 356 amino acid open reading frame. The central region of the predicted 38.5-kDa polypeptide of this clone was very similar to the protein product of the HSA stomatin gene EPB72 (Hiebl-Dirschmied et al., 1991; Wang et al., 1991; Stewart et al., 1992) and was named stomatin-like 2 (STOML2) and it is the second member, after SLP1 (Seidel et al., 1998), of the stomatin-like domain family to be identified in humans.

The STOML2 predicted polypeptide was shown to consist of three major domains; an N-terminal α -helical domain; a highly conserved central domain corresponding to ProDom domain PD001059 consisting of alternating α -helix and β -sheet and a C-terminal domain that was mostly α -helical. Because of its homology to stomatin the STOML2 polypeptide was predicted to be cytoplasmic. However, STOML2 did not have a transmembrane domain like other members of the mec-2 and Stomatin family of proteins. Also, the region in the sequence of the predicted STOML2 polypeptide, which aligns the closest with the mec-2 and STOMATIN transmembrane domains, shows a large content of charged and polar amino acids (see Fig. 4). It is therefore likely that STOML2 only exists in the intracellular compartment. These amino acids might form a short N-terminal domain in STOML2 instead of the transmembrane domain. Tavernarakis and Driscoll (1999) have recently described the SPFH domain superfamily, named for its core members: stomatins, prohibitins, flotillins and HflK/C. STOML2 is predicted to be a member of the stomatin branch of the SPFH domain superfamily. The biological function of most

of the SPFH domain family members is unknown although several genes in *C. elegans* have been identified: mec-2 (Huang et al., 1995), plays an essential role in mechanotransduction; unc-24 (Barnes et al., 1996) is required for normal locomotion; unc-1 (Rajaram et al., 1998) is also required for locomotion and interacts with genes involved in the response to volatile anesthetics. Interestingly, the mutations that disrupt these *C. elegans* genes are in conserved amino acid residues in the SPFH/stomatin-like domain indicating the functional importance of this region. The members of the SPFH domain superfamily appear to be multidomain proteins with a conserved SPFH domain and other domain(s) that specify the particular function of each protein (Fig. 6). The SPFH/stomatin-like domain is thought to tether the protein to the plasma membrane and may also have a regulatory function. Despite the absence of a transmembrane domain the existence of N-myristoylation sites suggest that STOML2 might still have a similar function as the other family members, however, its function might be subject to a quite different regulatory mechanism. One might speculate that the permanently membrane-bound SPFH family members behave more like receptors that get activated by ligand binding events. In contrast STOML2 might be activated only if it is recruited to the membrane surface by a myristoylation event, similar e.g. to Ras (Dunphy and Linder, 1998). A potential orthologue for STOML2 in *C. elegans* was hypothetical protein f3a010.5 (Fig. 5) however, to date no genetic mutants have been identified for this gene. The STOML2 gene is located on HSA chromosome 9p13. Although this region has been implicated in containing the gene responsible for acromesomelic dysplasia, Maroteaux Type (AMDM) (Kant et al., 1998) and

has been found to be rearranged in some cancers (Dave et al., 1995; Kim et al., 1997) it is unclear whether STOML2 could play a role in any of these disorders. (During the preparation of this manuscript an identical cDNA was described by Wang and Morrow, 2000).

Acknowledgments

The UK MRC HGMP is acknowledged for provision of radiation hybrid DNA samples and analysis programs.

References

- Barnes TM, Jin Y, Horvitz HR, Ruvkun G, Hekimi S: The *Caenorhabditis elegans* behavioral gene *Unc-24* encodes a novel bipartite protein similar to both erythrocyte band 7.2 (Stomatin) and nonspecific lipid transfer protein. *J Neurochem* 671:46–57 (1996).
- Braun BS, Frieden R, Lessnick SL, May WA, Denny CT: Identification of target genes for the Ewing's sarcoma EWS/FLI fusion protein by representational difference analysis. *Mol cell Biol* 15:4623–4630 (1995).
- Breathnach R, Chambon P: Organization and expression of eukaryotic split genes coding for proteins. *A Rev Biochem* 50:349–383 (1981).
- Corpet F, Gouzy J, Kahn D: Recent improvements of the ProDom database of protein domain families. *Nucl Acids Res* 27:263–267 (1999).
- Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton G: JPRED: A Consensus Secondary Structure Prediction Server. *Bioinformatics* 14:892–893 (1998).
- Dave BJ, Hopwood VL, King TM, Jiang H, Spitz MR, Pathak S: Genetic susceptibility to lung cancer as determined by lymphocytic chromosome analysis. *Cancer Epidemiol Biomarkers Prev* 4:743–749 (1995).
- Dunphy JT, Linder ME: Signalling functions of protein palmitoylation. *Biochim biophys Acta* 1436:245–261 (1998).
- Gonda TJ, Sheiness DK, Bishop JM: Transcripts from the cellular homologues of retroviral oncogenes: distribution among chicken tissues. *Mol Cell Biol* 2:617–624 (1982).
- Graw SL, Gardiner K, Hall-Johnson K, Hart I, Joeatham A, Walton K, Donaldson D, Patterson D: Molecular analysis and breakpoint definition of a set of human chromosome 21 somatic cell hybrids. *Somat Cell molec Genet* 21:415–428 (1995).
- Gress TM, Wallrapp C, Frohme M, Muller-Pillasch F, Lacher U, Friess H, Buchler M, Adler G, Hoheisel JD: Identification of genes with specific expression in pancreatic cancer by cDNA representational difference analysis. *Genes Chrom Cancer* 19:97–103 (1997).
- Gyapay G, Schmitt K, Fizes C, Jones H, Vega-Czarany N, Spillet D, Muselet D, Prud'Homme JF, Dib C, Auffray C, Morissette J, Weissenbach J, Goodfellow PN: A radiation hybrid map of the human genome. *Hum Mol Genet* 5:339–346 (1996).
- Hayes PD, Schmitt K, Jones HB, Gyapay G, Weissenbach J, Goodfellow PN: Regional assignment of human ESTs by whole-genome radiation hybrid mapping. *Mammal Genome* 7:446–450 (1996).
- Hertzog PJ, Hwang SY, Holland KA, Tymms MJ, Iannello R, Kola I: A gene on human chromosome 21 located in the region 21q22.2 to 21q22.3 encodes a factor necessary for signal transduction and antiviral response to type I interferons. *J Biol Chem* 269:19:14088–14093 (1994).
- Hiebl-Dirschmied C, Adolf GR, Prohaska R: Isolation and partial characterization of the human erythrocyte band 7 integral membrane protein. *Biochim biophys Acta* 1065:195–202 (1991).
- Hiebl-Dirschmied CM, Entler B, Glotzmann C, Maurer-Fogy I, Stratowa C, Prohaska R: Cloning and nucleotide sequence of cDNA encoding human erythrocyte band 7 integral membrane protein. *Biochim biophys Acta* 1090:123–124 (1991).
- Holland KA, Owczarek CM, Hwang SY, Tymms MJ, Constantinescu SN, Pfeiffer LM, Kola I, Hertzog PJ: A type I interferon signaling factor, ISF21, encoded on chromosome 21 is distinct from receptor components and their down-regulation and is necessary for transcriptional activation of interferon-regulated genes. *J Biol Chem* 272:21045–21051 (1997).
- Huang M, Gu G, Ferguson EL, Chalfie M: A stomatin-like protein necessary for mechanosensation in *C. elegans*. *Nature* 378:654:292–295 (1995).
- Hubank M, Schatz DG: Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucl Acids Res* 22:5640–5648 (1994).
- Jones DT: Do transmembrane protein superfolds exist? *FEBS Lett* 423:281–285 (1998).
- Jones DT: Protein secondary structure prediction based on position-specific scoring matrices. *J molec Biol* 292:195–202 (1999).
- Jones DT, Taylor WR, Thornton JM: A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 33:3038–3049 (1994).
- Kant SG, Polinkovsky A, Mundlos S, Zabel B, Thomeer RT, Zonderland HM, Shih L, van Haeringen A, Warman ML: Acromesomelic dysplasia Maroteaux type maps to human chromosome 9. *Am J hum Genet* 63:155–162 (1998).
- Kim SK, Ro JY, Kemp BL, Lee JS, Kwon TJ, Fong KM, Sekido Y, Minna JD, Hong WK, Mao L: Identification of three distinct tumor suppressor loci on the short arm of chromosome 9 in small cell lung cancer. *Cancer Res* 57:400–403 (1997).
- Kozak M: An analysis of the 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucl Acids Res* 15:8125–8147 (1987).
- Lewis BC, Shim H, Li Q, Wu CS, Lee LA, Maity A, Dang CV: Identification of putative c-Myc-responsive genes: characterization of rcl, a novel growth-related gene. *Mol Cell Biol* 17:4967–4978 (1997).
- Liu N, Lamerdin JE, Tucker JD, Zhou ZQ, Walter CA, Albala JS, Busch DB, Thompson LH: The human XRCC9 gene corrects chromosomal instability and mutagen sensitivities in CHO UV40 cells. *Proc natl Acad Sci, USA* 94:9232–9237 (1997).
- Niwa H, Harrison LC, DeAizpurua HJ, Cram DS: Identification of pancreatic beta cell-related genes by representational difference analysis. *Endocrinol* 138:1419–1426 (1997).
- Owczarek CM, Hwang SY, Holland KA, Gulluyan LM, Tavaría M, Weaver B, Reich NC, Kola I, Hertzog PJ: Cloning and characterization of soluble and transmembrane isoforms of a novel component of the murine type I interferon receptor, *Ifnar2*. *J Biol Chem* 272:23865–23870 (1997).
- Rajaram S, Sedensky MM, Morgan PG: Unc-1 – a stomatin homologue controls sensitivity to volatile anesthetics in *Caenorhabditis elegans*. *Proc natl Acad Sci, USA* 95:8761–8766 (1998).
- Raz R, Cheung K, Ling L, Levy DE: Three distinct loci on human chromosome 21 contribute to interferon-alpha/beta responsiveness. *Somat Cell mol Genet* 21:139–145 (1995).
- Salzer U, Ahorn H, Prohaska R: Identification of the phosphorylation site on human erythrocyte band 7 integral membrane protein: implications for a monotopic protein structure. *Biochim biophys Acta* 1151:149–152 (1993).
- Seidel G, Prohaska R: Molecular cloning of hSLP-1, a novel human brain-specific member of the band 7 MEC-2 family similar to *Caenorhabditis elegans* UNC-24. *Gene* 225:1–23–29 (1998).
- Shapiro MB, Senepathy P: RNA splice junctions of different classes: sequence statistics and functional implications in gene expression. *Nucl Acids Res* 15:7155–7174 (1987).
- Sharp PA: Speculations on RNA splicing. *Cell* 23:634–646 (1981).
- Snyers L, Umlauf E, Prohaska R: Oligomeric nature of the integral membrane protein stomatin. *J Biol Chem* 273:17221–17226 (1998).
- Stewart GW: Stomatin. *Int J Biochem Cell Biol* 29:271–274 (1997).
- Stewart GW, Argent AC, Dash BC: Stomatin: a putative cation transport regulator in the red cell membrane. *Biochim biophys Acta* 1225:15–25 (1993).
- Stewart GW, Hepworth-Jones BE, Keen JN, Dash BC, Argent AC, Casimir CM: Isolation of a cDNA coding for a ubiquitous membrane protein deficient in high Na⁺, low K⁺ stomatocytic erythrocytes. *Blood* 79:1593–1601 (1992).
- Tavernarakis N, Driscoll M: The SPFH domain: implicated in regulating targeted protein turnover in stomatins and other membrane-associated proteins. *Trend Biochem Sci* 24:425–427 (1999).
- Thompson JD, Higgins DG, Gibson TJ: CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res* 22:4673–4680 (1994).
- Wada J, Kumar A, Ota K, Wallner EI, Battle DC, Kanwar YS: Representational difference analysis of cDNA of genes expressed in embryonic kidney. *Kidney Internat* 51:1629–1638 (1997).
- Wang D, Mentzer WC, Cameron T, Johnson RM: Purification of band 7.2b, a 31-kDa integral phosphoprotein absent in hereditary stomatocytosis. *J Biol Chem* 266:27:17826–17831 (1991).
- Wang Y, Morrow JS: Identification and characterization of human SLP-2, a novel homologue of Stomatin (Band 7.2b) present in erythrocytes and other tissues. *J Biol Chem* 275:8062–8071 (2000).
- Zhu YW, Paszty C, Turetsky T, Tsai S, Kuypers FA, Lee G, Cooper P, Gallagher PG, Stevens ME, Rubin E, Mohandas N, Mentzer WC: Stomatocytosis is absent in "stomatin"-deficient murine red blood cells. *Blood* 93:2404–2410 (1999).

SeqServer[®]
biology in silico

ClustalW Results

Sequences

Help

Retrieval

BLAST2

FASTA

ClustalW

CGC Assembly

Phrap

Translation

Confidential -- Property of Incyte Corporation SeqServer Version 4.6 Jan 2002

☐ 789094CD1

☐ AF282596

CLUSTAL W (1.7) Multiple Sequence Alignments

Sequence format is Pearson

Sequence 1: 789094CD1 356 aa

Sequence 2: AF282596 356 aa

Start of Pairwise alignments

Aligning...

Sequences (1:2) Aligned. Score: 94

Start of Multiple Alignment

There are 1 groups

Aligning...

Group 1: Sequences: 2 Score:4464

Alignment Score 1975

CLUSTAL-Alignment file created [baaeqaGHf.aln]

CLUSTAL W (1.7) multiple sequence alignment

789094CD1 MLARAARGHWGPFAEG--LSTGFWPRSGRASSGLPRNTVVLFPQQEAWVVERMGRFHRI
AF282596 MLARAARGTGALLRGSLLASGRAPR--RASSGLPRNTVVLFPQQEAWVVERMGRFHRI

***** . : . * * : * ** *****

789094CD1 LEPGLNILIPVLDRIRYVQSLKEIVINVPEQSAVTLDNVTQLIDGVLYLRIMDPYKASYG
AF282596 LEPGLNILIPVLDRIRYVQSLKEIVINVPEQSAVTLDNVTQLIDGVLYLRIMDPYKASYG

789094CD1 VEDPEYAVTQLAQTTMRSELGKLSXDKVFRERESLNASIVDAINQAADCWGIRCLRYEIK
AF282596 VEDPEYAVTQLAQTTMRSELGKLSLXDKVFRERESLNASIVDAINQAADCWGIRCLRYEIK

789094CD1 DIHVPPRVKESMQMQVEAERRKRATVLESEGTRESAINVAEGKKQAQILASEAEKAEQIN
AF282596 DIHVPPRVKESMQMQVEAERRKRATVLESEGTRESAINVAEGKKQAQILASEAEKAEQIN

789094CD1 QAAGEASAVLAKAKAKAEAIRILAAALTQHNGDAAASLTVAEQYVSASFSLAKDSNTILL
AF282596 QAAGEASAVLAKAKAKAEAIRILAAALTQHNGDAAASLTVAEQYVSASFSLAKDSNTILL

789094CD1
AF282596

PSNPGDVTSMVAQAMGVYGALTKAPVPGTPDSLSSGSSRDVQGT
PSNPGDVTSMVAQAMGVYGALTKAPVPGTPDSLSSGSSRDVQGT

Submit sequences to:



ORIGINAL ARTICLE

Britta Fricke · Gordon W. Stewart ·
Kathryn J. Treharne · Anil Mehta · Gisela Knöpfle ·
Nicolaus Friedrichs · Klaus-Michael Müller ·
Monika von Düring

Stomatin immunoreactivity in ciliated cells of the human airway epithelium

Accepted: 18 March 2003 / Published online: 21 May 2003
© Springer-Verlag 2003

Abstract Stomatin is a widely distributed 32kD membrane protein of unknown function. In biochemical studies it is associated with cholesterol+sphingomyelin-rich 'rafts' in the cytomembrane. Genetic studies in *C. elegans*, supported by microscopic studies in mammalian tissue and co-expression studies in oocytes, suggest a functional link with the DEG/ENaC (degenerin/epithelial Na⁺ channel) superfamily of monovalent ion channels. Since ENaC channels play a prominent role in the physiology of the respiratory epithelium, we have studied the immunolocalization of stomatin in mature and developing human airway epithelium by means of Western blot analysis, immunocytochemistry, and immunoelectron microscopy. Stomatin immunoreactivity (stomatin-IR) was found in the ciliated cells of the conductive airway epithelium in a distinct distribution pattern with the strongest signal along the cilia. Immunogold labelling revealed immunogold particles at the basal bodies, along the cilia, and at the membrane of the microvilli. The

presence of stomatin-IR paralleled the stages of ciliogenesis in airway development, and its appearance preceded the elongation of the axoneme and the ciliary outgrowth. Due to its presence in the different cellular locations in the ciliated cell, we suggest that stomatin is involved in various cellular functions. From its ultrastructural position, stomatin could be a candidate for a membrane-associated mechanotransducer with a role in the control of ciliary motility. Stomatin as a raft protein might be a microtubule associated protein moving along the outer surface of the microtubules to its terminal site of action in the cilia. Stomatin-IR in microvilli supports the hypothesis of a co-localization with β - and γ -ENaC and, in conclusion, their potential functional interaction to control the composition of periciliary mucus electrolytes.

Keywords Stomatin · Cilia · Lung · Fetal · Mechanotransduction

B. Fricke (✉) · M. von Düring
Department of Neuroanatomy,
Institute of Anatomy MA 6/152, Ruhr University,
Universitätsstrasse 150, 44801 Bochum, Germany
e-mail: britta.fricke@ruhr-uni-bochum.de
Tel.: +49-234-3224768
Fax: +49-234-3214457

G. W. Stewart
Department of Medicine, University College London,
School of Medicine, Rayne Institute,
5 University Street, London, WC1E 6JJ, UK

K. J. Treharne · A. Mehta
Department of Child Health,
Ninewells Hospital and Medical School,
University of Dundee, Dundee, UK

G. Knöpfle · N. Friedrichs
Institute of Pathology, University of Bonn,
Sigmund-Freud-Strasse 25, 53127 Bonn, Germany

K.-M. Müller
Institute of Pathology, Bergmannsheil, Ruhr University,
Bürkle-de-la-Camp-Platz 1, 44789 Bochum, Germany

Introduction

Stomatin, or 'erythrocyte membrane protein 7.2b' (EMP72B), is a 32-kDa integral membrane protein that was first described in human erythrocytes (Hiebl-Dirschmied et al. 1991; Stewart et al. 1992). Attention was drawn to the protein by its absence from the red cell membrane in the dominantly-inherited haemolytic anaemia, hereditary stomatocytosis, in which the red cells show a catastrophic leak to the univalent cations Na⁺ and K⁺ (Lock et al. 1961; Lande et al. 1982; Eber et al. 1989; Stewart and Turner 1999; Stewart and Fricke 2003). However, the stomatin gene is not mutated in this condition, and this anaemia does not represent the phenotype of a human stomatin 'knock out' (Wang et al. 1992; Fricke et al. 2003). In these patients, the protein is present both in all of the other cells in which stomatin is found in normal human tissue (e.g. neutrophils, platelets), but also in the immature erythroid cells: it is lost from the red cell as it matures, possibly as a secondary result of the cation leak.

Since its original purification from erythrocytes, it has been recognised that the stomatin protein and its homologues are very widely distributed in prokaryotes and eukaryotes. A number of stomatin-like proteins are found in mammals. The function of this family in the various tissues and species is improperly understood (Rajaram et al. 1998, 1999; Sedensky et al. 2001; Tavernarakis et al. 1999; Boute et al. 2000; Gilles et al. 2000; Wang and Morrow 2000). Recent data have shown that stomatin is a 'raft' protein, partly associated with cholesterol+sphingomyelin-rich material in membranes (Snyers et al. 1999; Salzer et al. 2002; Salzer and Prohaska 2001; Mairhofer et al. 2002). There is a suggestion that the fundamental role of these proteins may lie in the control of surface expression of membrane proteins (Tavernarakis et al. 1999). Molecular biological data and immunocytochemical results in the rat, the mouse and the nematode *C. elegans* (*Caenorhabditis elegans*) give experimental evidence that the stomatin homologues might be directly involved in the process of mechanotransduction (Huang et al. 1995; Gu et al. 1996; Mannsfeldt et al. 1999; Hamill and Martinac 2001; Goodman et al. 2002). Mutations in the gene of the *C. elegans* homologue MEC-2, cause dysfunction in the process of mechanotransduction. The nematode mechanotransduction process also involves members of the DEG/ENaC (degenerin/epithelial Na⁺ channel) superfamily of ion channels, and stomatin is co-expressed with such channels in mammalian mechanosensory neurons (Tavernarakis and Driscoll 1997; Fricke et al. 2000). The DEG/ENaC superfamily of ion channels is named after the first described family members, the degenerins (DEG) in *C. elegans* and the mammalian epithelial amiloride-sensitive Na⁺ channel (ENaC) (Driscoll and Chalfie 1991; Canessa et al. 1993, 1994). Co-expression of the *C. elegans* stomatin homologue MEC-2 and the mutant degenerins MEC-4 and MEC-10 in *Xenopus* oocytes increased the activity of the mutant ion channels, supporting the suggestion that stomatin homologues might have a regulatory function controlling the activity of the associated ion channels, perhaps serving as a link to channel and cytoskeleton (Goodman et al. 2002).

The subunits of the epithelial amiloride-sensitive ion channels (ENaC) are abundantly expressed in pulmonary epithelia, where they are found in the ciliated cells of the epithelium of the conductive airways (Matsushita et al. 1996; Farman et al. 1997; Venkatesh and Katzberg 1997; Gaillard et al. 2000) and subserve an important role in the control of Na⁺ homeostasis and normal function of the lung (Hummler et al. 1997; Barker et al. 1998). Recently, Kobayakawa et al. (2002) revealed that stomatin-related olfactory protein (SRO)-mRNA is specifically expressed in murine olfactory receptor cells and that the protein is localized in their cilia. Up to now, the subcellular location of stomatin in the conductive airways is unknown. The present study was performed to investigate firstly, the stomatin-IR in ciliated cells of the airway epithelium, secondly, its subcellular location, and thirdly, the ontogenetic time course of its cellular location.

Materials and methods

Antisera

A rabbit polyclonal antibody was raised against recombinant stomatin fusion proteins (both hexa-histidine tagged and glutathione-S-transferase) containing amino acids 144-288(end) of the protein (Coles et al. 1999). The antibody was affinity-purified using a maltose-binding-protein-stomatin fusion protein. Although the affinity-purified antibody specifically recognizes in Western blots the expected 32-kDa band of human red cell stomatin, it could in theory recognize closely related family members that are not yet cloned. Therefore we will use the term stomatin immunoreactivity (stomatin-IR). To check for red cell contamination, western blots of nasal epithelial brushings were additionally stained with monoclonal antibodies against glycophorin A and glycophorin C+D (antibodies BRIC 163 and BGRL 100 respectively, both from the International Blood Group Reference laboratory, Bristol, U.K.).

Western blot

Nasal epithelial cells were obtained by a brushing technique (Treharne et al. 1994); sonicated in homogenisation buffer and subjected to centrifugation on a sucrose gradient (10-60% w/v), as described (Muimo et al. 2000). Fractions were harvested and subjected to SDS-PAGE, western blotted, and the blot probed with antibodies directed against stomatin (1:10,000), glycophorin A (1:5), glycophorin C +D (1:5). Antibody binding was detected by the ECL system (Amersham, U.K.). Previous studies using iodinated ciliary cells have shown that in this experiment, the cilia are typically found in the lower part of the gradient at >60% sucrose w/v (Treharne and Mehta, unpublished), i.e. lanes 1-6. Apical membranes are typically found at around 30-40% sucrose (lanes 7-9).

Tissue preparation

Archived paraffin-embedded tissue of routine bronchiolar biopsies of 8 patients was used to analyse the tissue from adults. For developmental studies, archived post-mortem paraffin-embedded embryonic, fetal and neonatal tissue (ten patients ranging from 7 to 37 weeks gestational age, 6 days and 2 months) was analysed. All fetuses were well preserved without any signs of maceration. The Ethical Review Board on Use of Human Subjects in Research of the University of Bonn (Nr. 177/00) approved these experiments.

Immunocytochemistry

The tissue was fixed in a 4% formaldehyde solution and embedded in paraffin. Paraffin sections, 7 µm in thickness, were prepared. Prior to the incubation, antigen retrieval was performed (following the laboratory recommendations of Dianova, Hamburg, Germany). The sections were rehydrated, rinsed in deionized water and incubated in a target retrieval solution (10 mM citrate buffer, pH 6) for 15 min using a microwave oven (800 W). The specimens were washed in phosphate-buffered saline (PBS) and preincubated in a solution containing 20% normal goat serum (NGS), 0.1% Triton X-100 and 0.05% phenylhydrazine in PBS for 30 min at room temperature. The primary polyclonal antibody (anti-stomatin), diluted 1:500 and 1:100 in a solution consisting of 20% NGS, 0.1% Triton X-100, 0.01% thimerosal and 0.1% sodium azide in PBS, was applied for 16 h at room temperature in a humidified chamber. The specimens were then exposed to biotinylated goat anti-rabbit IgG (Vector Laboratories, Burlingame, Calif., USA), diluted 1:2000 in PBS-A (2 mg/ml bovine serum albumin in PBS, 0.1% sodium azide w/v) for 4 h at room temperature. For antigen visualization, the sections were incubated with the ABC reagent (avidin-biotinylated-peroxidase complex; Vector Laboratories, Burlingame, Calif., USA) for 2 h at room temperature, diluted

1:1000 in PBS-A. The substrate reaction was performed with 3,3'-diaminobenzidine. The sections were counterstained with hematoxylin, dehydrated in graded ethanol and mounted with EntellanR (Merck). For negative controls, the primary antibody was omitted.

Electron microscopy

Bronchial biopsies from patients were subsequently prepared for transmission electron microscopy postembedding immunogold labelling by fixation in 2.5% buffered glutaraldehyde and postfixation with osmium tetroxide, dehydrated in graded ethanol and finally embedded in Epon.

For immunogold labelling, 75 nm ultrathin sections were prepared. Postembedding immunogold labelling was performed by a modification of the protocol of Mahendrasingam et al. (1997). Before incubation, the sections were pretreated with 5% H₂O₂ in distilled water for 5 min, then washed in TBS (Tris-buffered saline) and preincubated in 5% donkey serum in TBS for 30 min. The primary polyclonal antibody (anti-stomatin), diluted 1:100 in 1% TBS/BSA (bovine serum albumin), was applied for 20 h at 4°C in a humidified chamber. The specimens were washed in 1% BSA in TBS, preincubated with 5% donkey serum in TBS for 30 min and exposed to the 18 nm colloidal gold-AffiniPure donkey anti-rabbit IgG (Jackson Immunological, West Grove, Phil., USA), diluted 1:10 in 1% BSA/TBS. After washing in TBS and distilled water, the sections were counterstained with 2% uranylacetate in 70% methanol followed by lead citrate. The specimens were analysed in a Philips electron microscope 400. For the negative control, the primary antibody was omitted. Red cells in the tissue served as positive controls for correct antibody staining.

Results

Stomatin-IR in ciliated cells of adult airway epithelium

Western blot

Figure 1 shows western blots of the sucrose-gradient-fractionated human respiratory epithelium, obtained from fresh nasal brush biopsies and probed with the affinity-purified stomatin antibody and red-cell-specific anti-glycophorin antibodies. The anti-stomatin antibody (Panel A), used at 1 in 10,000 dilution, revealed a single major band at 32 kDa, the expected molecular mass of human stomatin (Stewart et al. 1992). The intensity was maximal at 40% w/v sucrose. Previous studies on animal epithelium which was surface iodinated with lactoperoxidase (Treharne and Mehta, unpublished) have shown that in such sucrose gradient experiments, the distal parts of the cilia are found in the lower part of the gradient at 60% sucrose w/v, i.e. lanes 1-4, while apical membranes, also containing the basal bodies of the cilia, are typically found at around 40-55% sucrose (lanes 5-8). The anti-glycophorin A and anti-glycophorin C+D antibodies, both used at 1 in 5 dilution, each revealed dual peaks at about 50% sucrose (lanes 2, 3) and 30% (lanes 7, 8), different from the main stomatin immunoreactivity at 40%, suggesting that the stomatin signal in this brushings preparation is non-red-cell in origin. In any case, since these latter two antibodies are normally used in this laboratory to detect the proteins in red cell preparations at

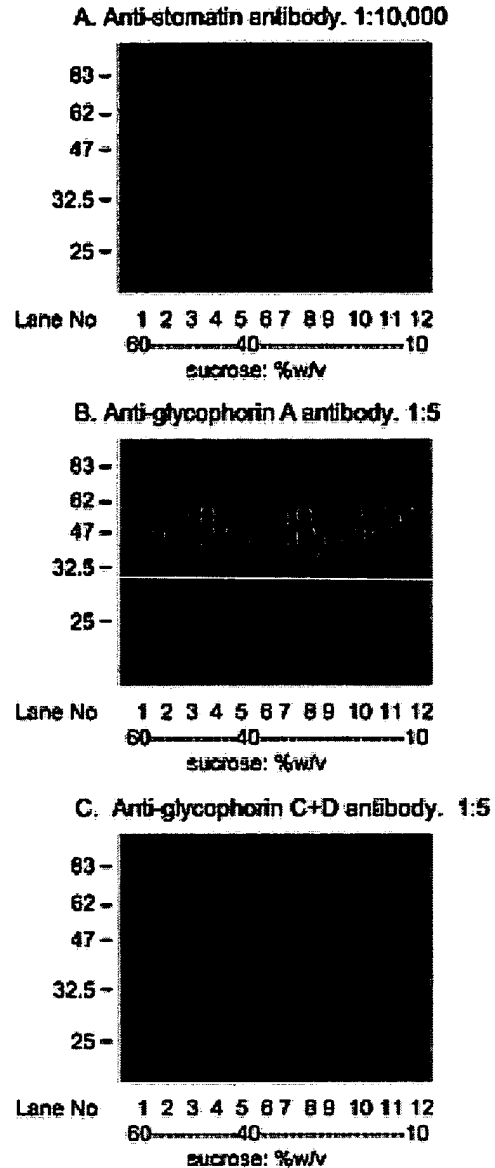


Fig. 1A-C Western blotting of human airway epithelial brushings. Cells were obtained by brushing nasal epithelium and subjected to sucrose gradient centrifugation. The fractions were run on SDS gels and blotted to nitrocellulose. The nitrocellulose was probed with anti-stomatin (A), anti-glycophorin A (B) and anti-glycophorin B+C (C) antibodies. The latter two are specific for the red cell membrane. While the anti-stomatin antibody was used at a dilution of 1 in 10,000, the anti-glycophorin antibodies had to be used at 1 in 5, reflecting the low red cell contamination of the sample. The stomatin was most evident at 40% (w/v) sucrose while all of the glycophorins were enriched in two maxima, at about 50% sucrose and 25% sucrose respectively, showing that red cell membranes migrated at different sucrose densities from the predominant stomatin band at 40% sucrose

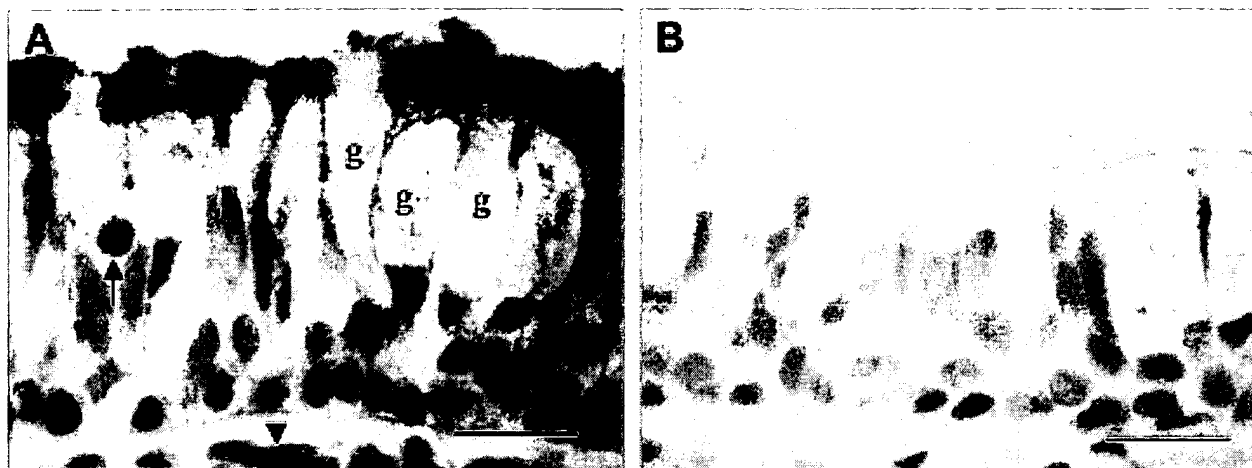


Fig. 2 Adult bronchial pseudostratified epithelium. **A** Stomatin-IR is concentrated in the cilia of the airway epithelium. Goblet cells (g) do not exhibit stomatin-IR. Invaded stomatin-positive lympho-

cyte (arrow) and macrophage (arrowhead). **B** Negative control. Bar 20 μ m

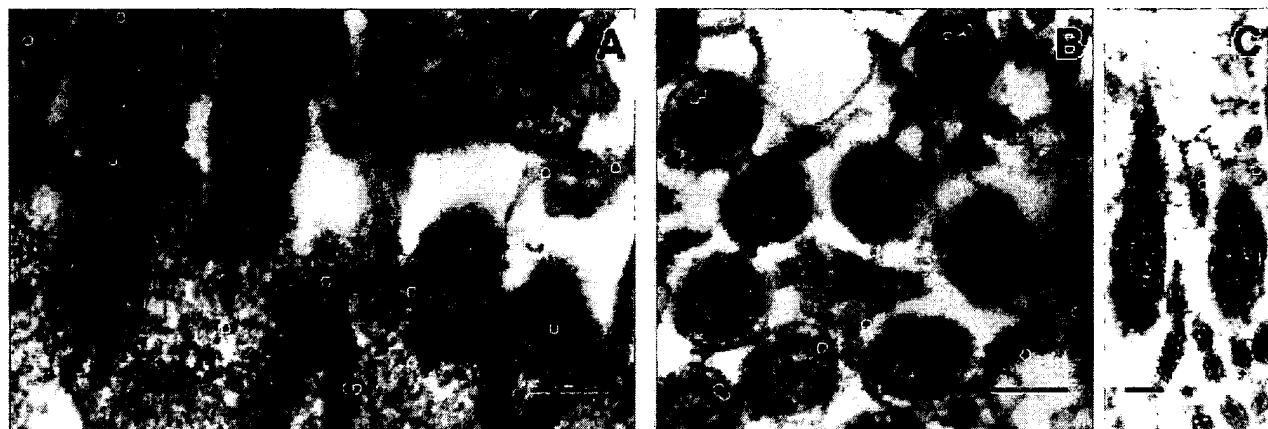


Fig. 3A–C Ultrathin sections of pseudostratified bronchial epithelium, immunogold-labelling using the anti-stomatin antibody. Immunogold particles are regularly located along the axoneme associated to the outer microtubule doublets and at the basal bodies.

A Note the location of the immunogold particles at the membrane and the microtubules of the cilia in **B** and at the membrane of the microvilli in **C**. Bar 200 nm

titres of $>1:500$, the quantity of red cell contamination is minimal.

Light microscopy

Stomatin-IR was detected in ciliated cells of all segments of the conductive airways up to the cuboidal ciliated cells in the respiratory bronchioles (Fig. 2A). Immunoreactivity was also seen in the apical portion of Clara cells in bronchioles. Neither goblet cells, nor the alveolar epithelial cells (type I and type II) exhibited stomatin-IR. In the ciliated cells, stomatin-IR was most intense in the apical portion of the cell where basal bodies, cilia and microvilli are located (Fig. 2). Tangential sections of the apical cytoplasm clearly showed a dotted line of immunoreaction product consistent with the labelling of the basal

bodies. In the normal adult pseudostratified epithelium, basal cells did not show stomatin-IR. We also found invaded stomatin-immunopositive lymphocytes, which are known to express stomatin (Fricke et al. 2003).

Immunoelectron microscopy

On the ultrastructural level, immunogold particles were found in three main locations in the ciliated cells: first, along the cilia, where immunogold particles were regularly located between the membrane and the outer microtubule doublets of the axoneme; second, at the basal bodies in close association with the microtubules; and third, at the membrane of the microvilli (Fig. 3). No immunogold particles were found in the cytoplasm of the goblet cells.

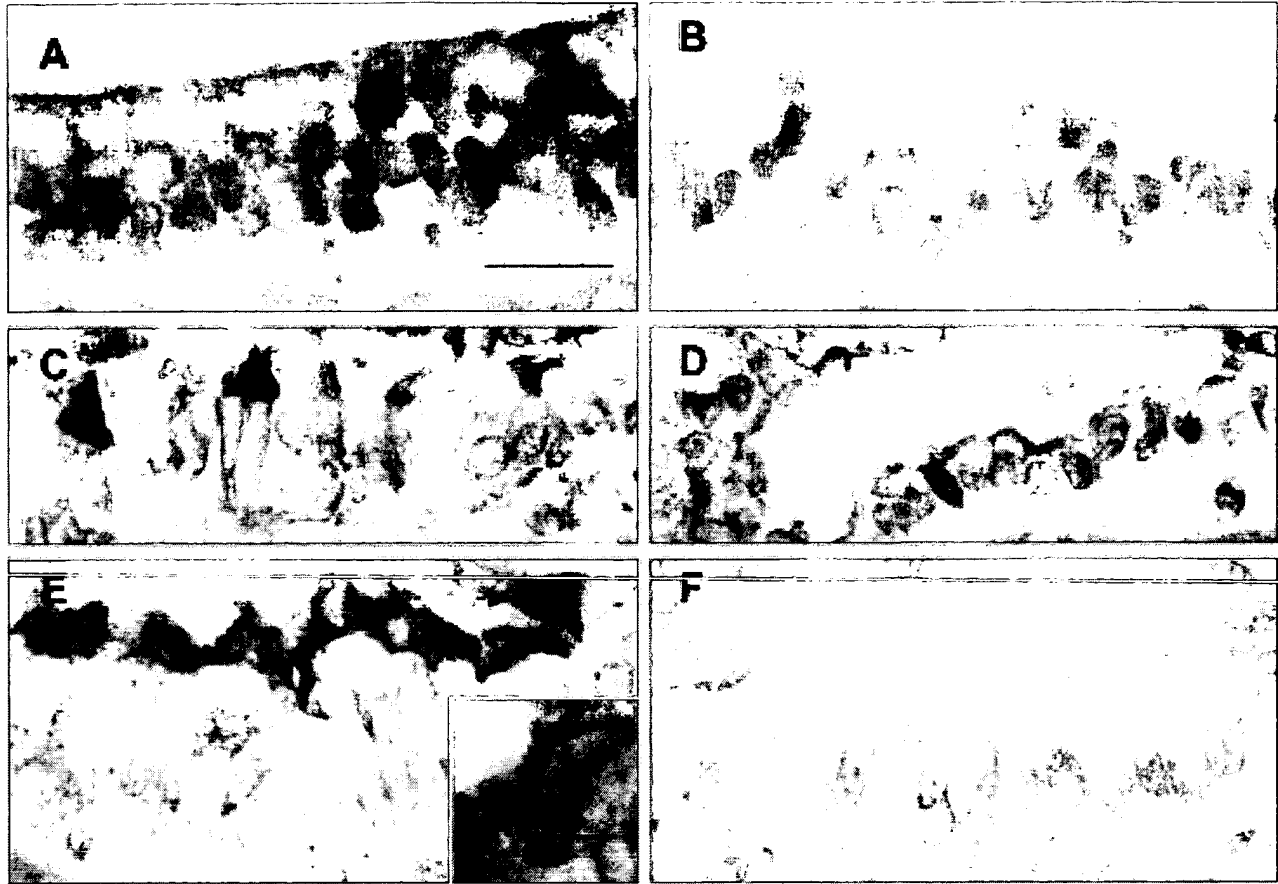


Fig. 4A–F Ontogenetic expression of stomatin-IR in ciliated epithelium of the conductive airways. *Bar* 20 μ m. **A, B** Early expression of stomatin-IR in the trachea at gestational age of 7 weeks, before development of the cilia. **A** Immunopositive staining is accentuated close to the apical plasma membrane. Basal cells exhibit stomatin immunoreactivity; **B** Negative control. **C, D** Pseudoglandular period. **C** Strong immunoreactivity in the cilia of

the airway epithelium of the bronchial segment. **D** In distal segments of the bronchial tree stomatin-IR is located in the apical membrane of the cells. **E, F** Canalicular period. **E** Stomatin-immunopositive pseudostratified bronchial epithelium. Note the dotted line of immunoreaction product consistent with the labelling of the basal bodies (*inset*, *bar* 5 μ m). **F** Negative control

Stomatin-IR precedes elongation of cilia during ciliogenesis

At a gestational age of 7 weeks, in which ciliogenesis begins in the membranous trachea, an immunopositive staining of the apical cytoplasm of the tracheal epithelium was observed (Fig. 4A). The immunoreactivity became accentuated close to the apical plasma membrane. In the undifferentiated pseudostratified epithelium, all basal cells exhibited stomatin-IR, which was lost during maturation and further differentiation.

In the pseudoglandular period (14 weeks of gestation) and canalicular stage (19 weeks of gestation), during which ciliogenesis becomes obvious in the more distal segments of the bronchial tree, the ciliated epithelium of the proximal bronchi showed a strong staining in the apical region with the most intense immunoreactivity in the cilia itself (Fig. 4C). Tangential sections of the apical compartment clearly showed an accumulation of stomatin-IR in the subplasmalemmal region where the basal

bodies become differentiated (Fig. 4E). In the more distal segments of the conductive airways the cells are in the process of ciliogenesis, showing differentiation of basal bodies (kinetosomes) and the differentiation of the peripheral ciliary apparatus. During these developmental steps of ciliogenesis, stomatin-IR was enhanced in the apical portion of the cells (Fig. 4D).

Furthermore, we observed stomatin-IR in the embryonic and fetal connective tissue cells. The analysis of the various developmental stages revealed that stomatin-IR was pronounced in the developing lung connective tissue starting in the embryonic period, but reduced later during maturation of the lung and almost absent in the alveolar phase.

Discussion

The results described in this paper clearly demonstrate that stomatin-IR is strongly exhibited in ciliated cells of

the adult human airway epithelium from a very early developmental stage. During ciliogenesis, the presence of the immunoreaction product was first observed in the subplasmalemmal region where the basal bodies become differentiated. In the mature airway epithelium, both light microscopy and immunogold labelling revealed the strongest immunoreactivity at the basal bodies and along the cilia, in which the immunoreaction product was located between the membrane and the outer microtubules, and at the membrane of microvilli. Stomatin-IR was absent from goblet cells. Immunoreactivity was also seen in the endothelium of blood vessels and in the peripheral cells of the blood, as expected from previous work (Stewart and Fricke 2003; Fricke et al. 2003).

Comparison of the present data with published data on the distribution of ENaC proteins (Gaillard et al. 2000) shows that stomatin-IR is found at every location where β - and γ -ENaC are found, especially the apical cell region of ciliated cells and in the microvilli. However, the distribution of stomatin-IR extends beyond that of β - and γ -ENaC, being found in the cilia, where β - and γ -ENaC are not located. The presence of stomatin-IR also precedes the expression of ENaC in development: while stomatin is present at 7 weeks, ENaC is not evident until much later, at the canalicular stage (after 17 weeks). It remains possible that ENaC function depends on stomatin, an assertion that is not inconsistent with previous observations in rodents and *C. elegans* (Huang et al. 1995; Gu et al. 1996; Goodman et al. 2002).

Although the structural organization of the 15-protofilament microtubules in *C. elegans* mechanosensory neurons is different from the axonemal microtubules in cilia, the close association of microtubules, plasma membrane and extracellular matrix is a striking similarity between the cilia of the airway epithelium, where stomatin-IR is particularly localized, and the mechanosensory neurons of *C. elegans*, which express the stomatin homologue MEC-2 and whose function is abrogated in *mec-2* mutations (Huang and Chalfie 1994; Huang et al. 1995; Driscoll and Kaplan 1997). In the *C. elegans* model, MEC-2 is discussed as a linker molecule between plasma membrane, cytoskeleton, and membrane-associated ion channels (Huang et al. 1995; Gu et al. 1996; Hamill and Martinac 2001; Goodman et al. 2002). The localization of stomatin in the cilia of the airway epithelium between plasma membrane and the outer microtubule doublets of the axoneme is consistent with this hypothesis.

Could the role of stomatin in cilia be related to the mechanism of mechanotransduction? Up to now it remains to be elucidated in cilia of the airway epithelium, how beating frequency and synchronization of the beat mechanism is controlled. Mechanotransduction could be important in cilia for two physiological reasons: one, for position detection during beating; and two, for detection of fluid flow over the cilia, causing deflection of the cilium. The primary cilium of mammalian renal epithelia is mechanically sensitive and is discussed to serve as a flow sensor (Praetorius and Spring 2001). Positioned between plasma membrane and microtubules of the cilia,

stomatin in the ciliated respiratory cell could be a candidate for a membrane-associated mechanotransducer involved in the control of ciliary motility.

It is striking that in early developmental stages of lung development, stomatin is not only expressed in the undifferentiated pseudostratified epithelium, but also in the connective tissue cells. The fact that stomatin is lost in these cells during the maturation process strongly underlines the time dependent expression of stomatin during ontogenesis. During development of the respiratory system, mechanical stress is exerted on the airway epithelium as well as on the connective tissue cells. This mechanical stress seems to be an important regulatory mechanism for proper fetal lung cell proliferation and differentiation (Liu et al. 1999; Liu and Post 2000). As a membrane-associated protein early expressed during ciliogenesis, stomatin might have a potential role in the pathophysiology of primary ciliary dyskinesia (PCD).

Acknowledgements We thank the Sir Jules Thorn Trust for funding (GWS). We thank Luzie Augustinowski, Margaret Chetty and Katja Rumpf for their excellent technical assistance, Debbie Baines and Monica Driscoll for useful discussions.

References

- Barker PM, Nguyen MS, Gatzky JT, Grubb B, Norman H, Hummler E, Rossier B, Boucher RC, Koller B (1998) Role of γ -ENaC subunit in lung liquid clearance and electrolyte balance in newborn mice. Insights into perinatal adaptation and pseudo-hypoaldosteronism. *J Clin Invest* 102:1634–1640
- Boute N, Gribouval O, Roselli S, Benessy F, Lee H, Fuchshuber A, Dahan K, Gubler MC, Niaudet P, Antignac C (2000) NPHS2, encoding the glomerular protein podocin, is mutated in autosomal recessive steroid-resistant nephrotic syndrome. *Nat Genet* 24:349–354
- Canessa CM, Horisberger JD, Rossier BC (1993) Epithelial sodium channel related to proteins involved in neurodegeneration. *Nature* 361:467–470
- Canessa CM, Schild L, Buell G, Thorens B, Gautschi I, Horisberger JD, Rossier BC (1994) Amiloride-sensitive epithelial Na⁺ channel is made of three homologous subunits. *Nature* 367:463–466
- Coles SE, Ho MM, Chetty MC, Nicolaou A, Stewart GW (1999) Hereditary stomatocytosis with marked pseudohyperkalemia. *Br J Haematol* 104:275–283
- Driscoll M, Chalfie M (1991) The *mec-4* gene is a member of a family of *Caenorhabditis elegans* genes that can mutate to induce neuronal degeneration. *Nature* 349:588–593
- Driscoll M, Kaplan J (1997). Mechanotransduction. In: Riddle DL, Blumenthal T, Meyer BJ, Priess JR (eds) *The nematode C. elegans*, II. Cold Spring Harbor Press, Cold Spring Harbor, New York, pp 645–677
- Eber SW, Lande WM, Iarocci TA, Mentzer WC, Hohn P, Wiley JS, Schroter W (1989) Hereditary stomatocytosis: consistent association with an integral membrane protein deficiency. *Br J Haem* 72:452–455
- Farman N, Talbot CR, Boucher R, Fay M, Canessa C, Rossier BC, Bonvalet JP (1997) Noncoordinated expression of α -, β -, γ -subunit mRNAs of epithelial Na⁺ channel along the rat respiratory tract. *Am J Physiol* 272:131–141
- Fricke B, Argent AC, Pizzey AR, Chetty MC, Turner EJ, Ho MM, Jolascon A, Düring M von, Stewart GW (2003) The 'stomatin' gene and protein in overhydrated hereditary stomatocytosis. *Blood* (in press)

- Fricke B, Lints R, Stewart GW, Drummond H, Dodt G, Driscoll M, Düring M von (2000) Epithelial Na⁺ channels and stomatin are expressed in rat trigeminal mechanosensory neurons. *Cell Tissue Res* 299:327–334
- Gaillard D, Hinnrasky J, Coscoy S, Hofman P, Matthay MA, Puchelle E, Barbry P (2000) Early expression of β - and γ -gamma-subunits of epithelial sodium channel during human airway development. *Am J Physiol Lung Cell Mol Physiol* 278:177–184
- Gilles F, Glenn M, Goy A, Remache Y, Zelenetz AD (2000) A novel gene STORP (STOmatin-Related Protein) is localized 2 kb upstream of the promyelocytic gene on chromosome 15q22. *Eur J Haematol* 64:104–113
- Goodman MB, Ernstrom GG, Chelur DS, ÓHagan R, Yao CA, Chalfie M (2002) MEC-2 regulates *C. elegans* DEG/ENAC channels needed for mechanosensation. *Nature* 415:1039–1042
- Gu G, Caldwell GA, Chalfie M (1996) Genetic interaction affecting touch sensitivity in *Caenorhabditis elegans*. *Proc Natl Acad Sci USA* 93:6577–6582
- Hamill OP, Martinac B (2001) Molecular basis of mechanotransduction in living cells. *Physiol Rev* 81:685–740
- Hiebl-Dirschmied CM, Adolf GR, Prohaska R (1991) Isolation and partial characterization of the human erythrocyte band 7 integral membrane protein. *Biochim Biophys Acta* 1065:195–202
- Huang M, Chalfie M (1994) Gene interactions affecting mechanosensory transduction in *Caenorhabditis elegans*. *Nature* 367:467–470
- Huang M, Gu G, Ferguson EL, Chalfie M (1995) A stomatin-like protein necessary for mechanosensation in *C. elegans*. *Nature* 378:292–295
- Hummeler E, Barker P, Talbot C, Wang Q, Verdunno C, Grubb B, Gatzky J, Burnier M, Horisberger JD, Beermann F, Boucher RC, Rossier BC (1997) A mouse model for the renal salt-wasting syndrome pseudohypoaldosteronism. *Proc Natl Acad Sci USA* 94:11710–11715
- Kobayakawa K, Hayashi R, Morita K, Miyamichi K, Oka Y, Tsuboi A, Sakano H (2002) Stomatin-related olfactory protein, SRO, specifically expressed in the murine olfactory sensory neurons. *J Neurosci* 22:5931–5937
- Lande WM, Thiemann PW, Mentzer WC (1982) Missing band 7 membrane protein in two patients with high sodium, low potassium red cells. *J Clin Invest* 70:1273–1280
- Liu M, Post M (2000) Mechanical signal transduction in the fetal lung. *J Appl Physiol* 89:2078–2084
- Liu M, Tanswell K, Post M (1999) Mechanical force-induced signal transduction in lung cells. *Am J Physiol* 277:667–683
- Lock SP, Sephton Smith R, Hardisty RM (1961) Stomatocytosis: a hereditary haemolytic anomaly associated with haemolytic anemia. *Br J Haematol* 7:303–314
- Mahendrasingam S, Katori Y, Furness DN, Hackney CM (1997) Ultrastructural localization of cadherin in the adult guinea-pig organ of Corti. *Hear Res* 111:85–92
- Mairhofer M, Steiner M, Mosgoeller W, Prohaska R, Salzer U (2002) Stomatin is a major lipid-raft component of platelet alpha granules. *Blood* 100:897–904
- Mannsfeldt AG, Stucky CP, Lewin GR (1999) Stomatin, a MEC-2 like protein, is expressed by mammalian sensory neurons. *Mol Cell Neurosci* 13:391–404
- Matsushita K, McCray PB, Sigmund RD, Welsh MJ, Stokes JB (1996) Localization of epithelium sodium channel subunit mRNAs in adult rat lung by in situ hybridization. *Am J Physiol* 271:332–339
- Muimo R, Hornickova Z, Riemen CE, Gerke V, Matthews H, Mehta A (2000) Histidine phosphorylation of annexin I in airway epithelia. *J Biol Chem* 275:36632–36636
- Praetorius HA, Spring KR (2001) Bending the MDCK cell primary cilium increases intracellular calcium. *J Membr Biol* 184:71–79
- Rajaram SR, Sedensky MM, Morgan PG (1998) Unc-1: a stomatin homologue controls sensitivity to volatile anesthetics in *Caenorhabditis elegans*. *Proc Natl Acad Sci USA*, 95:8761–8766
- Rajaram SR, Spangler TL, Sedensky MM, Morgan PG (1999) A stomatin and a degenerin interact to control anesthetic sensitivity in *Caenorhabditis elegans*. *Genetics* 153:1673–1682
- Salzer U, Hinterdorfer P, Hunger U, Borken C, Prohaska R (2002) Ca(++)-dependent vesicle release from erythrocytes involves stomatin-specific lipid rafts, synexin (annexin VII), and sorcin. *Blood* 99:2569–2577
- Salzer U, Prohaska R (2001) Stomatin, flotillin-1, and flotillin-2 are major integral proteins of erythrocyte lipid rafts. *Blood* 97:1141–1143
- Sedensky MM, Siefker JM, Morgan PG (2001) Model organisms: new insights into ion channel and transporter function. Stomatin homologues interact in *Caenorhabditis elegans*. *J Physiol Cell Physiol* 280:1340–1348
- Snyers L, Umlauf E, Prohaska R (1999) Association of stomatin with lipid-protein complexes in the plasma membrane and the endocytotic compartment. *Eur J Cell Biol* 78:802–812
- Stewart GW, Hepworth-Jones BE, Keen JN, Dash BCJ, Argent AC, Casimir CM (1992) Isolation of cDNA coding for an ubiquitous membrane protein deficient in high Na⁺ low K⁺ stomatocytic erythrocytes. *Blood* 79:1593–1601
- Stewart GW, Fricke B (2003) The curious genomic path from leaky red cell to nephrotic kidney. *Nephron Physiol* 93:29–33
- Stewart GW, Turner EJ (1999) The hereditary stomatocytoses and allied disorders: congenital disorders of erythrocyte membrane permeability to Na⁺ and K⁺. *Baillieres Best Pract Res Clin Haematol* 12:707–728
- Tavernarakis N, Driscoll M (1997) Molecular modeling of mechanotransduction in the nematode *Caenorhabditis elegans*. *Annu Rev Physiol* 59:659–689
- Tavernarakis N, Driscoll M, Kyrpidis NC (1999) The SPFH domain: a universal motif in stomatins and other membrane-associated proteins implicated in regulating targeted protein turnover. *Trends Biochem Sci* 24:425–427
- Treharne KJ, Marshall LJ, Mehta A (1994) A novel chloride-dependent GTP-utilizing protein kinase in plasma membranes from human respiratory epithelium. *Am J Physiol* 267:592–601
- Venkatesh VC, Katzberg HD (1997) Glucocorticoid regulation of epithelial sodium channel genes in human fetal lung. *Am J Physiol Lung Cell Mol Physiol* 273:227–233
- Wang Y, Morrow JS (2000) Identification and characterization of human SLP-2, a novel homologue of stomatin (band 7.2b) present in erythrocytes and other tissues. *J Biol Chem* 275:8062–8071
- Wang D, Turetsky T, Perrine S, Johnson RM, Mentzer WC (1992) Further studies on RBC membrane protein 7.2B deficiency in hereditary stomatocytosis. *Blood* 80 [Suppl 1]:275

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and †Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests have evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6\$2.00/0
PNAS is available online at <http://www.pnas.org>.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "In-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith-Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it

is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or $\approx 0.5\%$ of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from <http://sss.stanford.edu/sss/>, and databases derived from the current version of SCOP may be found at <http://scop.mrc-lmb.cam.ac.uk/scop/>.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith-Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties $-12/-1$ (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

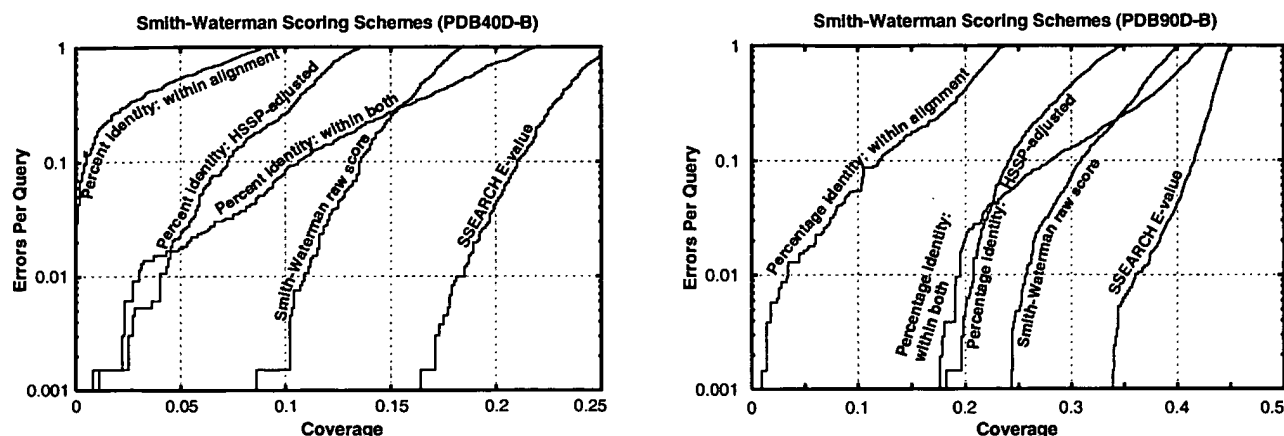


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith-Waterman. (A) Analysis of PDB40D-B database. (B) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H . Smith-Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how

protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Receiver Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

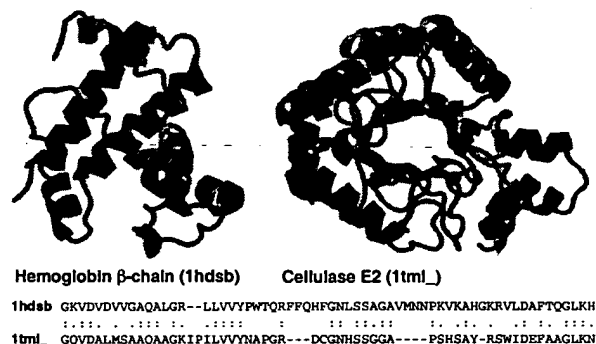


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β -chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASVOL (40).

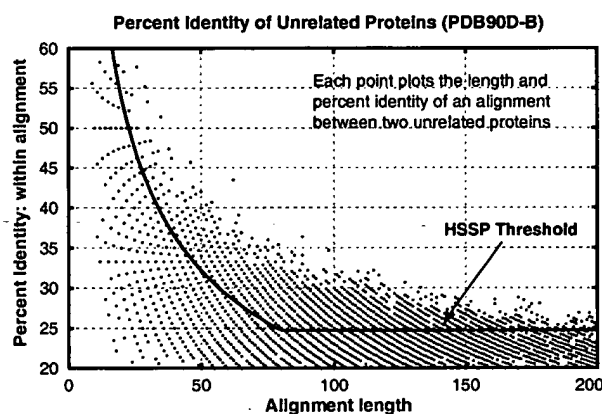


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

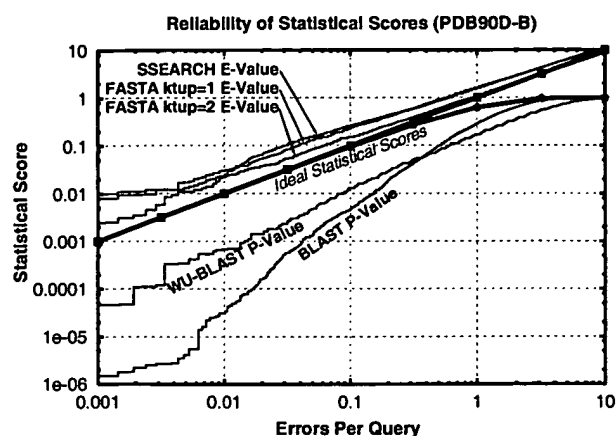


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure

related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith-Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

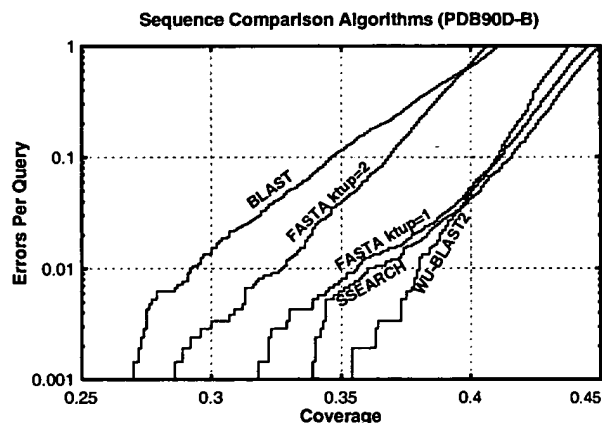
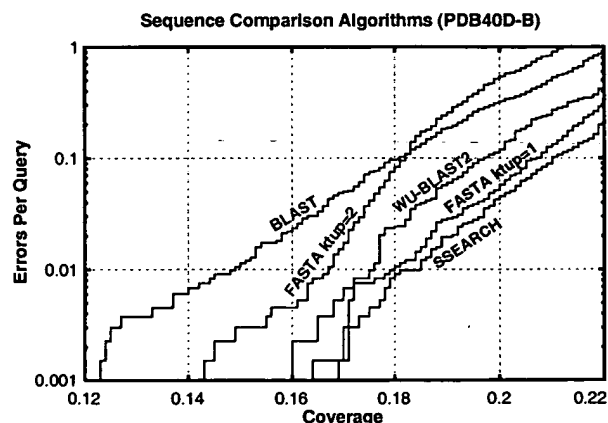


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (A) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (B) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA $k_{\text{up}} = 1$ is nearly as effective as SSEARCH. FASTA $k_{\text{up}} = 2$ and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA $k_{\text{up}} = 1$. WU-BLAST2 is slightly faster than FASTA $k_{\text{up}} = 2$, but the latter has more interpretable scores.

In PDB40D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA $k_{\text{up}} = 1$, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

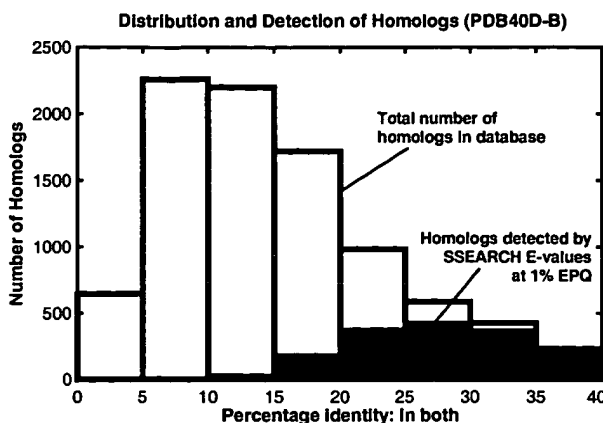


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA $k_{\text{up}} = 1$ E-values	3.9	0.03	17.9
FASTA $k_{\text{up}} = 2$ E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at <http://sss.stanford.edu/sss/>.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266**, 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266**, 635–643.
6. Pearson, W. R. (1991) *Genomics* **11**, 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4**, 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266**, 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249**, 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17**, 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24**, 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24**, 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Bio-medical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233**, 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1**, 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1**, 77–78.
21. Arratia, R., Gordon, L. & M. W. (1986) *Ann. Stat.* **14**, 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266**, 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9**, 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7**, 369–376.
32. Orengo, C., Michie, A., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. (1997) *Structure (London)* **5**, 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39**, 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20**, 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4**, 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131**, 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32**, 9906–9916.
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376.

Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE‡
and G. GORDON GIBSON*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey,
Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.

2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.

3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.

4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wynford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Hegningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

* Author for correspondence; e-mail: g.gibson@surrey.ac.uk

† Current Address: US Environmental Protection Agency, National Health and Environmental Effects, Research Laboratory, Reproductive Toxicology Division, Research Triangle Park, NC 27711, USA.

‡ Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.

altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinauer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

- (1) Differential screening,
- (2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
- (3) Differential display (DD),
- (4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
- (5) Gene expression arrays, and
- (6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences), EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original approaches used to identify such genes was described 20 years ago by St John and Davis (1979). These authors developed a method, termed 'differential plaque filter

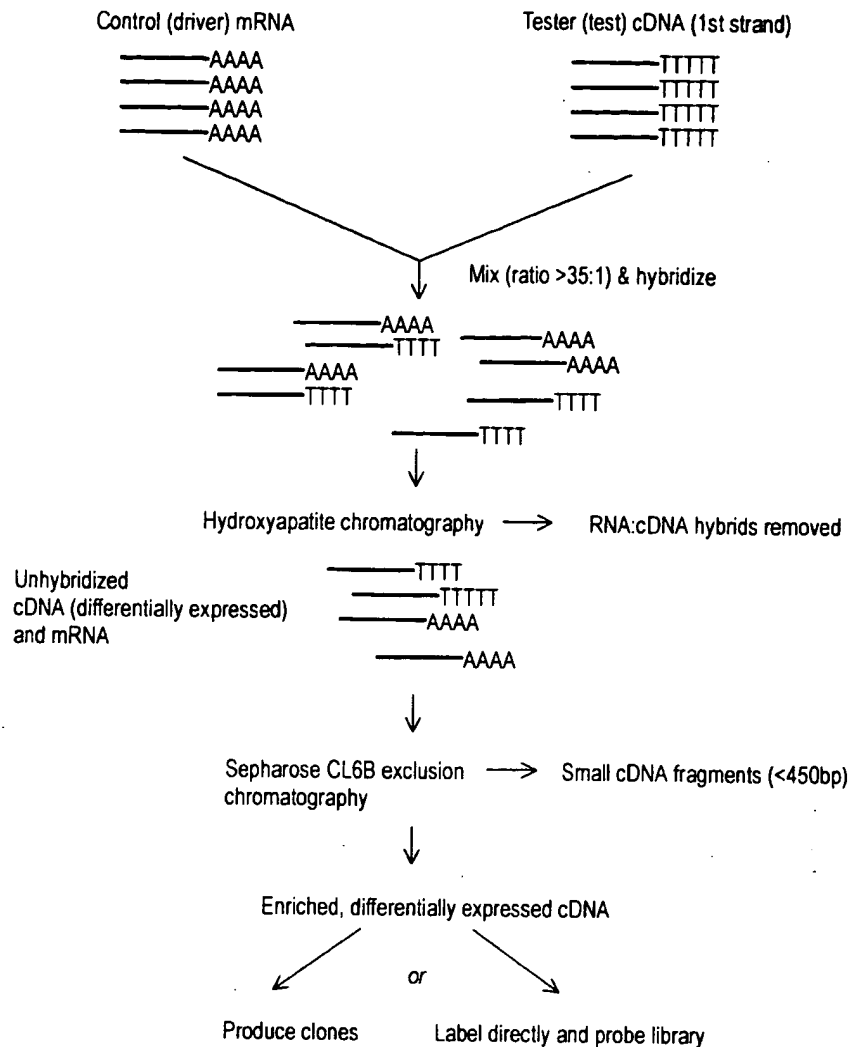


Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/alterd (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/alterd population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983).

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligo-vector and reduce the amplification potential of the control population. The digested control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column (streptavidin may also be used) to remove the control population, including heteroduplexes formed by annealing of common sequences from the tester population. The procedure is repeated several times following the addition of fresh

control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT₃₀) primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT₃₀ forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT₃₀ population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

Chemical Cross-Linking Subtraction (CCLS)

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridinyl-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 µg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random oligonucleotide primed (DROP) adaptation (Hampson *et al.* 1996, Hampson and Hampson 1997) uses random hexanucleotide sequences to prime solid phase-synthesized cDNA. Since each primer includes a T7 polymerase promotor sequence

at the 5' end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

Representational Difference Analysis (RDA)

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5' adaptors at each end of the DNA duplex and can, thus, be filled in at both 3' ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Schatz used a tester:driver ratio of 1:400, 1:80 000 and 1:800 000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1% of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplification and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-purified adaptors. These are free of the truncated adaptors which appear to be a major source of the false positive bands. A very similar technique to RDA, termed linker capture subtraction (LCS) was described by Yang and Sytowski (1996).

Suppression PCR Subtractive Hybridization (SSH)

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added to two portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complementary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene expression profiles of enzyme-inducers such as phenobarbital (Rockett *et al.* 1997) and Wy-14,643 (Rockett *et al.* unpublished observations). The isolation of differentially expressed genes in this manner enables the construction of a fingerprint

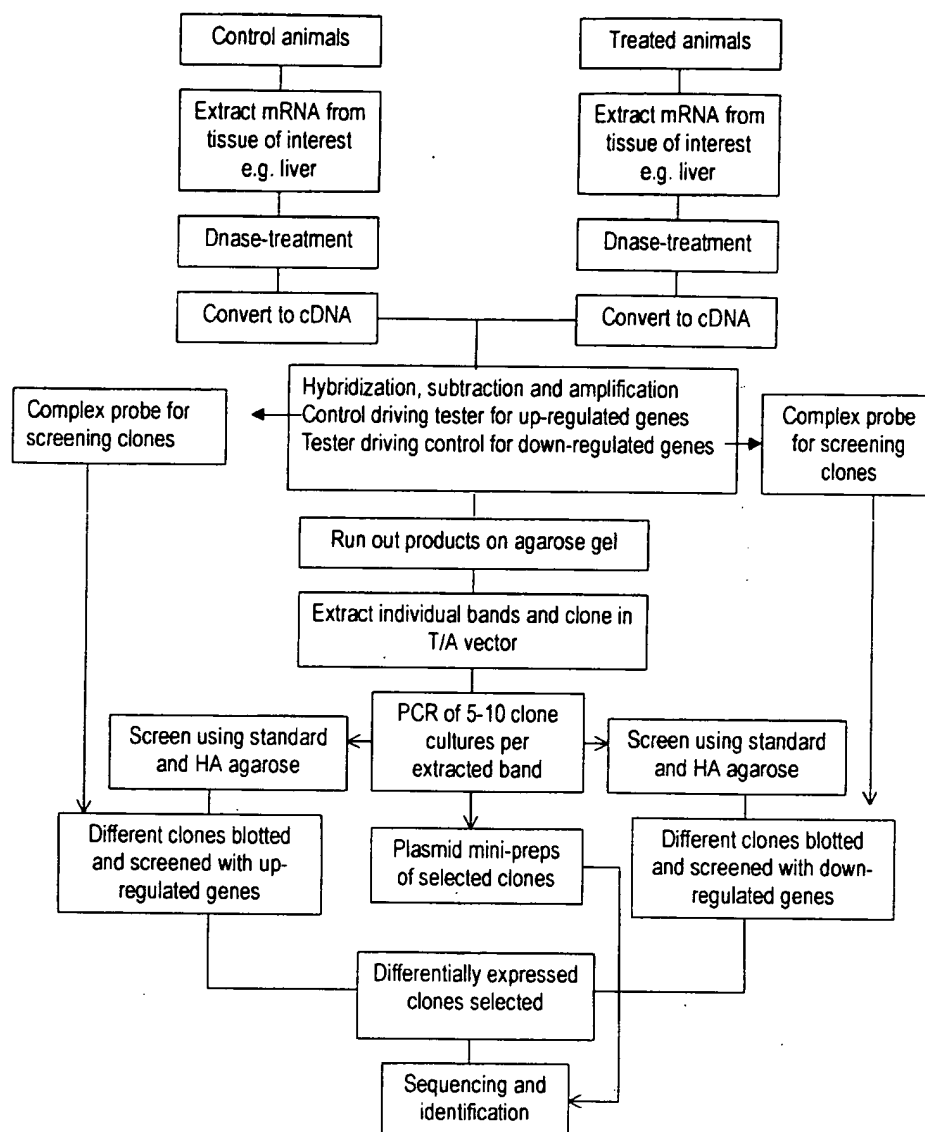


Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3).

One of the advantages in using the SSH approach is that no prior knowledge is required of which specific genes are up/down-regulated subsequent to xenobiotic

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
5 (1300)	93.5%	CYP2B1
7 (1000)	95.1%	Preproalbumin
		Serum albumin mRNA
8 (950)	98.3%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
10 (850)	95.7%	CYP2B1
11 (800)	Clone 1 94.9%	CYP2B1
	Clone 2 75.3%	CYP2B2
12 (750)	93.8%	TRPM-2 mRNA
		Sulfated glycoprotein
15 (600)	92.9%	Preproalbumin
		Serum albumin mRNA
16 (55)	Clone 1 95.2%	CYP2B1
	Clone 2 93.6%	Haptoglobin mRNA partial alpha
21 (350)	99.3%	18S, 5.8S & 28S rRNA

Bands 1-4, 6, 9, 13, 14, and 17-20 are shown to be false positives by dot blot analysis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
1 (1500)	95.3%	3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Hemopoxin mRNA
3 (1000)	91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1 77.2%	<i>M. musculus</i> C1 inhibitor
	Clone 2 94.5%	Electron transfer flavoprotein
	Clone 3 91.0%	<i>M. musculus</i> Topoisomerase 1 (Topo 1)
8 (650)	Clone 1 86.9%	Soares 2NbMT <i>M. musculus</i> (EST)
	Clone 2 96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1 86.9%	Soares mouse NML <i>M. musculus</i> (EST)
	Clone 2 82.0%	Soares p3NMF 19.5 <i>M. musculus</i> (EST)
10 (550)	73.8%	Soares mouse NML <i>M. musculus</i> (EST)
11 (525)	95.7%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
12 (375)	100.0%	Ribosomal protein
13 (23)	Clone 1 97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2 100.0%	Fibrinogen B-beta-chain
	Clone 3 100.0%	Apolipoprotein E gene
14 (170)	96.0%	Soares p3NMF19.5 <i>M. musculus</i> (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	<i>R. norvegicus</i> RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4-6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR (RT-PCR). The products are then run side-by-side on sequencing gels. Those bands which are present in one display only, or which are much more intense in one

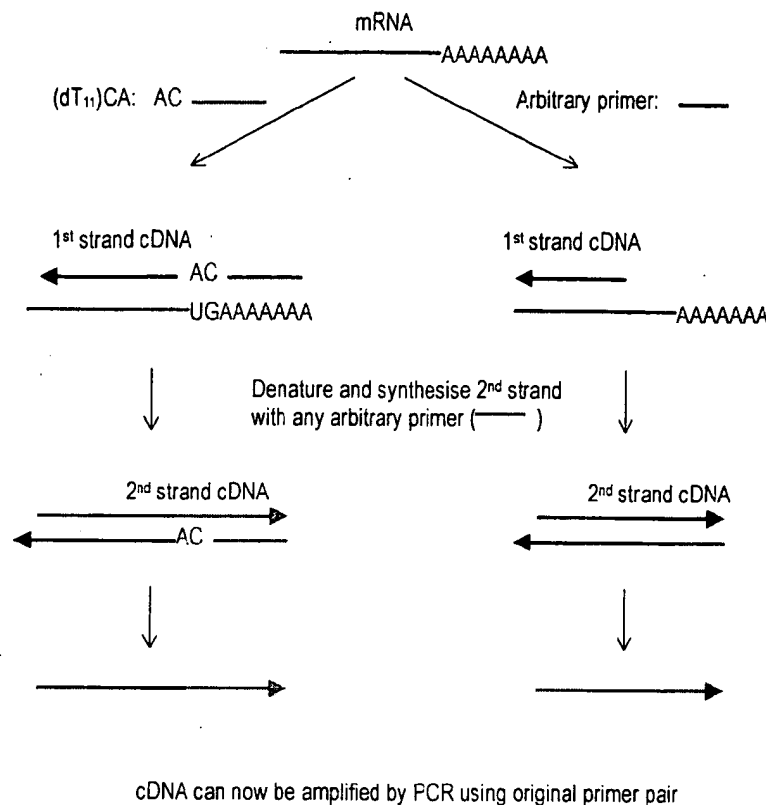


Figure 8. Two approaches to differential display (DD) analysis. 1st strand synthesis can be carried out either with a polydT₁₁NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polydT primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1st strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2nd strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2nd strand may also hybridize to the 1st strand cDNA in a number of different places, several different 2nd strand products may be obtained from one binding point of the 1st strand primer. Following 2nd strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

Restriction endonuclease-facilitated analysis of gene expression

Serial Analysis of Gene Expression (SAGE)

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatenation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this procedure, double stranded cDNA from the test cells is synthesized with a biotinylated polydT primer. Following digestion with a commonly cutting (4bp recognition sequence) restriction enzyme ('anchoring enzyme'), the 3' ends of the cDNA population are captured with streptavidin beads. The captured population is

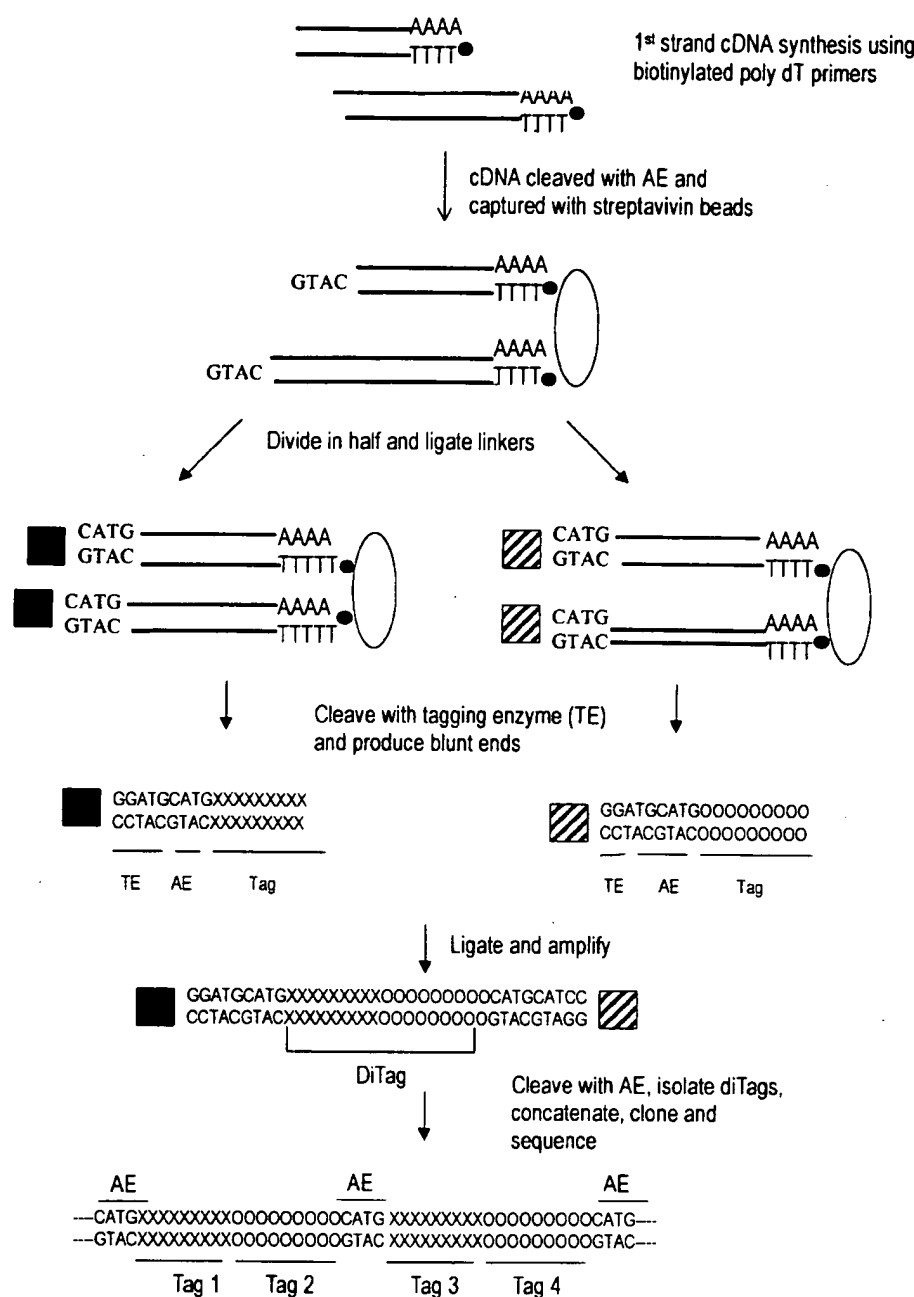


Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3' ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the ditags isolated from the linkers using PAGE. The ditags are then ligated (during which process, concatenation occurs) and cloned into a vector of choice for sequencing. After Velculescu *et al.* (1995), with permission.

of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmataz *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at <http://www.tigr.org>) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BAC)s, or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmataz *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

Problems and potential of differential expression techniques

The holistic or single cell approach?

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and/or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes/neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or *in situ* RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new microdissection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site: <http://www.ncbi.nlm.nih.gov/ncicgap/intro.html>). There are also separation techniques available that utilise cell-specific antigens as a means to isolate target cells,

species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10000× smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon- γ -stimulated HeLa cells differentially express up to 433 genes (assuming 24 000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38 000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate; Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3' restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockett, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

Whilst the latest differential display technologies are purported to include design and experimental modifications to overcome this lack of efficiency (in both the total number of differentially expressed genes recovered and the percentage that are true

experiments and animals. DD, on the other hand, is not subject to this grey zone since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

Resolution and visualization of differential expression products

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaPor HR (National Diagnostics, Hesse, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hanse Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wishes to check whether all the clones produced from a specific band in a differential display experiment are derived from the same gene species, a small amount of reamplified or digested clone can be run on a standard high resolution gel, and a second aliquot

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1995) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30%) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, overstaining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to over stain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

The possible use of 'microfingerprinting' to reduce complexity

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'micro-fingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000 + bp on the same gel, which leads to compromise in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to include certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different cellular effects. If the prognosis for exposure to one or more other chemicals which display a similar profile is already known, then one could perhaps predict similar effects for any new compounds which show a similar micro-fingerprint.

which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

- (1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
- (2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identity between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Although the procedure is somewhat more complex than Northern analysis, requiring synthesis of primers and optimization of reaction conditions for each gene species, it is now possible to set up high throughput PCR systems using multichannel pipettes, 96 +well plates and

become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

Conclusions

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5–10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One persistent problem is understanding whether differentially expressed genes are a cause or consequence of the altered state. Furthermore, many chemicals, such as non-genotoxic carcinogens, are also mitogens and so genes associated with replication will also be upregulated but may have little or nothing to do with the

US Environmental Protection Agency and approved for publication. Approval does not signify that the contents reflect the views and policies of the Agency, nor does mention of trade names constitute endorsement or recommendation for use.

References

- ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMERPOPOULOS, M. H., XIAO, H., MERRIL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., McCOMBIE, W. R. and VENTOR, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651-1656.
- AN, G., LUO, G., VELTRI, R. W. and O'HARA, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, **20**, 342-346.
- AXEL, R., FEIGELSON, P. and SCHULTZ, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7**, 247-254.
- BAND, V. and SAGER, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the National Academy of Sciences, USA*, **86**, 1249-1253.
- BAUER, D., MULLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOF, P. and STRAUSS, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272-4280.
- BERTIOLI, D. J., SCHLICHTER, U. H. A., ADAMS, M. J., BURROWS, P. R., STEINBISS, H.-H. and ANTONIW, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520-4523.
- BRAVO, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, **1**, 37-46.
- BURN, T. C., PETROVICK, M. S., HOHAUS, S., ROLLINS, B. J. and TENEN, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776-2783.
- CAO, J., CAI, X., ZHENG, L., GENG, L., SHI, Z., PAO, C. C. and ZHENG, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447-451.
- CASSIDY, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13-20.
- CHANG, G. W. and TERZAGHI-HOWE, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445-4452.
- CHEN, J., SCHWARTZ, D. A., YOUNG, T. A., NORRIS, J. S. and YAGER, J. D., 1996, Identification of genes whose expression is altered during mitosuppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, **17**, 2783-2786.
- CHEN, J. J. W. and PECK, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, **24**, 793-794.
- CLON TECHNIQUES, 1997a, PCR-Select Differential Screening Kit—the nextstep after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18-19.
- CLON TECHNIQUES, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTechniques*, **XII**, 15-16.
- DAVIS, M. M., COHEN, D. I., NIELSEN, E. A., STEINMETZ, M., PAUL, W. E. and HOOD, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha. *Proceedings of the National Academy of Sciences (USA)*, **81**, 2194-2198.
- DELLAVALLE, R. P., PETERSON, R. and LINDQUIST, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in *Drosophila melanogaster*. *Molecular and Cell Biology*, **14**, 3646-3659.
- DERISI, J. L., VASHWANATH, R. L. and BROWN, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.
- DIATCHENKO, L., LAU, Y.-F. C., CAMPBELL, A. P., CHENCHIK, A., MOQADAM, F., HUANG, B., LUKYANOV, K., GURSKAYA, N., SVERDLOV, E. D. and SIEBERT, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, **93**, 6025-6030.
- DOGRA, S. C., WHITELAW, M. L. and MAY, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1-9.
- DUGUID, J. R. and DINAUER, M. C., 1990, Library subtraction of *in vitro* cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research*, **18**, 2789-2792.
- DUNBAR, P. R., OGG, G. S., CHEN, J., RUST, N., VAN DER BRUGGEN, P. and CERUNDOLO, V., 1998, Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Current Biology*, **26**, 413-416.

- quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, **5**, 622-626.
- KILTY, I. and VICKERS, P., 1997, Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter* (Stratagene), **10**, 50-51.
- KLEINJAN, D.-J. and VAN HEYNINGEN, V., 1998, Position effect in human genetic disease. *Human and Molecular Genetics*, **7**, 1611-1618.
- KO, M. S., 1990, An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, **18**, 5705-5711.
- LAKE, B. G., EVANS, J. G., CUNNINGHAME, M. E. and PRICE, R. J., 1993, Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, **101**, 241-248.
- LAKE, B. G., EVANS, J. G., GRAY, T. J. B., KOROSI, S. A. and NORTH, C. J., 1989, Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guinea pig and marmoset. *Toxicology and Applied Pharmacology*, **99**, 148-160.
- LENNARD, M. S., 1993, Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, **9**, 60-77.
- LEVY, S., TODD, S. C. and MAECKER, H. T., 1998, CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, **16**, 89-109.
- LIANG, P. and PARDEE, A. B., 1992, Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967-971.
- LIANG, P., AVERBOUKH, L., KEYOMARSI, K., SAGER, R. and PARDEE, A., 1992, Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, **52**, 6966-6968.
- LIANG, P., AVERBOUKH, L. and PARDEE, A. B., 1993, Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, **21**, 3269-3275.
- LIANG, P., BAUER, D., AVERBOUKH, L., WARTHOF, P., ROHRWILD, M., MULLER, H., STRAUSS, M. and PARDEE, A. B., 1995, Analysis of altered gene expression by differential display. *Methods in Enzymology*, **254**, 304-321.
- LINSKENS, M. H., FENG, J., ANDREWS, W. H., ENLOW, B. E., SAATI, S. M., TONKIN, L. A., FUNK, W. D. and VILLEPONTEAU, B., 1995, Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, **23**, 3244-3251.
- LISITSYN, N., LISITSYN, N. and WIGLER, M., 1993, Cloning the differences between two complex genomes. *Science*, **259**, 946-951.
- LOHMANN, J., SCHICKLE, H. and BOSCH, T. C. G., 1995, REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, **18**, 200-202.
- LUNNEY, J. K., 1998, Cytokines orchestrating the immune response. *Reviews in Science and Technology*, **17**, 84-94.
- MAKOWSKA, J. M., GIBSON, G. G. and BONNER, F. W., 1992, Species differences in ciprofibrate-induction of hepatic cytochrome P450A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, **7**, 183-191.
- MALDARELLI, F., XIANG, C., CHAMOUN, G. and ZEICHNER, S. L., 1998, The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, **53**, 39-51.
- MATHIEU-DAUDE, F., CHENG, R., WELSH, J. and MCCLELLAND, M., 1996, Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, **24**, 1504-1507.
- MCKENZIE, D. and DRAKE, D., 1997, Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter* (Stratagene), **10**, 19-20.
- MCCLELLAND, M., MATHIEU-DAUDE, F. and WELSH, J., 1996, RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, **11**, 242-246.
- MECHLER, B. and RABBITTS, T. H., 1981, Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, **88**, 29-36.
- MEYER, U. A. and ZANGER, U. M., 1997, Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, **37**, 269-296.
- MOHLER, K. M. and BUTLER, L. D., 1991, Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, **28**, 437-447.
- MURPHY, L. D., HERZOG, C. E., RUDICK, J. B., TITO FOJO, A. and BATES, S. E., 1990, Use of the polymerase chain reaction in the quantitation of the *mdr-1* gene expression. *Biochemistry*, **29**, 10351-10356.
- NELSON, D. R., KOYMANS, L., KAMATAKI, T., STEGEMAN, J. J., FEYEREISEN, R., WAXMAN, D. J., WATERMAN, M. R., GOTOH, O., COON, M. J., ESTABROOK, R. W., GUNSALUS, I. C. and NEBERT, D. W., 1996, Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, **6**, 1-42.

- SUNG, Y. J. and DENMAN, R. B., 1997, Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, **23**, 462-464.
- SUTTON, G., WHITE, O., ADAMS, M. and KERLAVAGE, A., 1995, TIGR Assembler; A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, **1**, 9-19.
- SUZUKI, Y., SEKIYA, T. and HAYASHI, K., 1991, Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, **192**, 82-84.
- SYED, V., GU, W. and HECHT, N. B., 1997, Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, **18**, 264-273.
- UITTERLINDEN, A. G., SLAGBOOM, P., KNOOK, D. L. and VIJGL, J., 1989, Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (USA)*, **86**, 2742-2746.
- ULLMAN, K. S., NORTHROP, J. P., VERWEIJ, C. L. and CRABTREE, G. R., 1990, Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, **8**, 421-452.
- VASMATZIS, G., ESSAND, M., BRINKMANN, U., LEE, B. and PASTON, I., 1998, Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (USA)*, **95**, 300-304.
- VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K. W., 1995, Serial analysis of gene expression. *Science*, **270**, 484-487.
- VOELTZ, G. K. and STEITZ, J. A., 1998, AuuuA sequences direct mRNA deadenylation uncoupled from decay during *Xenopus* early development. *Molecular and Cell Biology*, **18**, 7537-7545.
- VOGELSTEIN, B. and KINZLER, K. W., 1993, The multistep nature of cancer. *Trends in Genetics*, **9**, 138-141.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13-14.
- WAN, J. S., SHARP, S. J., POIRIER, G. M.-C., WAGAMAN, P. C., CHAMBERS, J., PYATI, J., HOM, Y.-L., GALINDO, J. E., HUVAR, A., PETERSON, P. A., JACKSON, M. R. and ERLANDER, M. G., 1996, Cloning differentially expressed mRNAs. *Nature Biotechnology*, **14**, 1685-1691.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13-14.
- WANG, Z. and BROWN, D. D., 1991, A gene expression screen. *Proceedings of the National Academy of Sciences (USA)*, **88**, 11505-11509.
- WAWER, C., RUGGEBERG, H., MEYER, G. and MUYZER, G., 1995, A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, **23**, 4928-4929.
- WELSH, J., CHADA, K., DALAL, S. S., CHENG, R., RALPH, D. and MCCLELLAND, M., 1992, Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, **20**, 4965-4970.
- WONG, H., ANDERSON, W. D., CHENG, T. and RIABOWOL, K. T., 1994, Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, **223**, 251-258.
- WONG, K. K. and MCCLELLAND, M., 1994, Stress-inducible gene of *Salmonella typhimurium* identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (USA)*, **91**, 639-643.
- WYNFORD-THOMAS, D., 1991, Oncogenes and anti-oncogenes; the molecular basis of tumour behaviour. *Journal of Pathology*, **165**, 187-201.
- XHU, D., CHAN, W. L., LEUNG, B. P., HUANG, F. P., WHEELER, R., PIEDRAFITA, D., ROBINSON, J. H. and LIEW, F. Y., 1998, Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, **187**, 787-794.
- YANG, M. and SYTOWSKI, A. J., 1996, Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, **237**, 109-114.
- ZHAO, N., HASHIDA, H., TAKAHASHI, N., MISUMI, Y. and SAKAKI, Y., 1995, High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, **156**, 207-213.
- ZHAO, X. J., NEWSOME, J. T. and CIHLAR, R. L., 1998, Up-regulation of two *Candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, **25**, 121-129.
- ZIMMERMANN, C. R., ORR, W. C., LECLERC, R. F., BARNARD, C. and TIMBERLAKE, W. E., 1980, Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, **21**, 709-715.



Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene-expression. Such gene expression regulations account for both the

pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

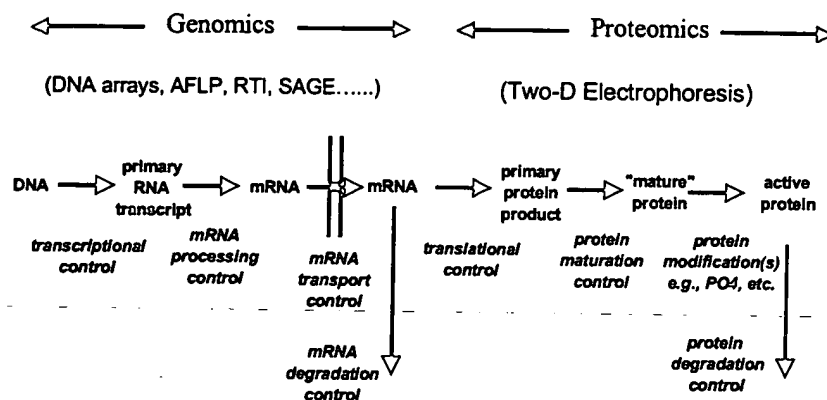


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

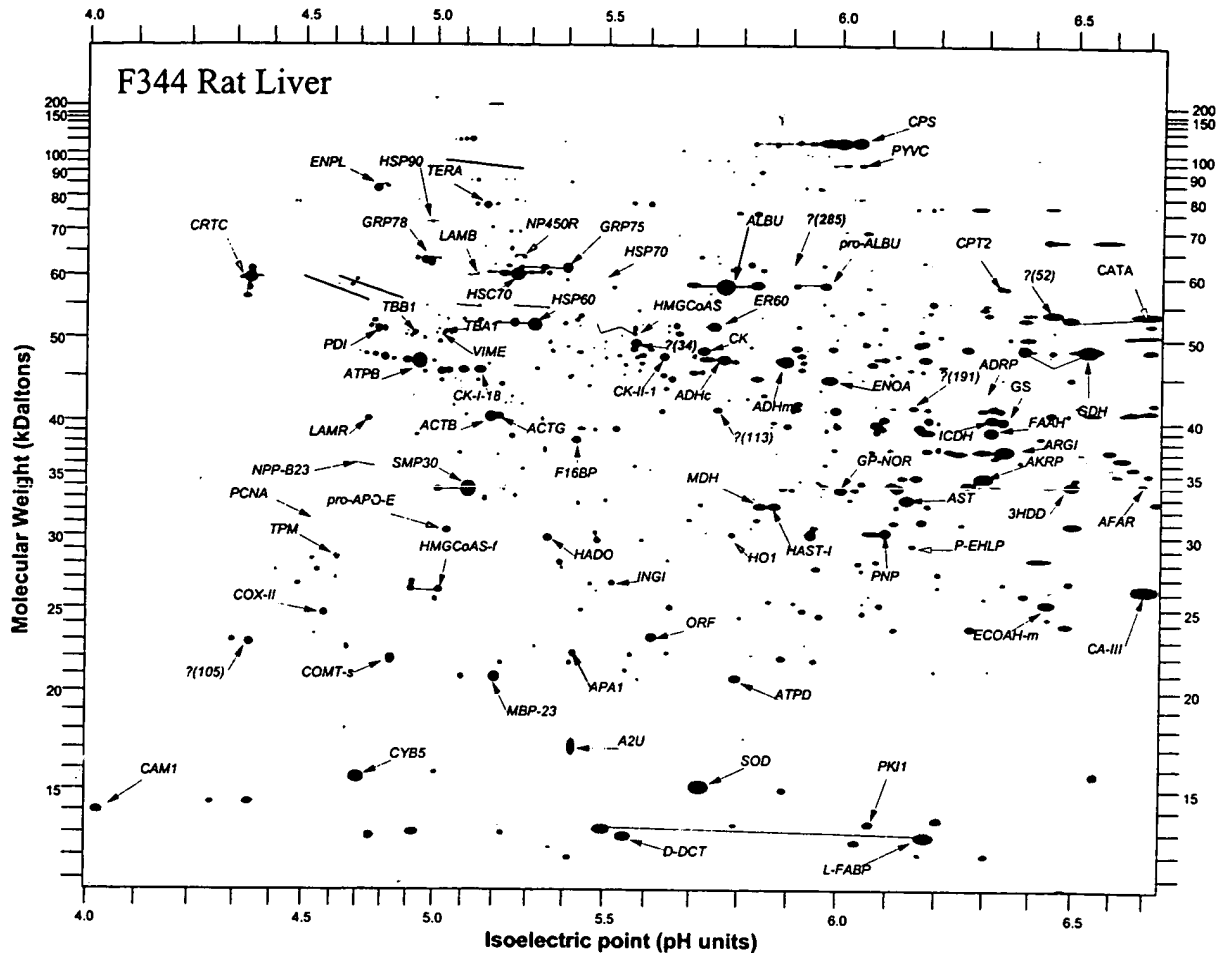


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplifi-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug devel-

opment, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trials.

References

- Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* 19, 1998–2003.
- Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537.
- Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* 12, 907–930.
- Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. *J. Cell. Biochem. Suppl.* 22, 108–116.
- Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137, 75–89.
- Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* 63, 2243–2250.

- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. *Science* 274, 610–614.
- Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. *Electrophoresis* 19, 355–363.
- Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773.
- Mann, M., Hojrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338–345.
- Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. *Electrophoresis* 14, 157–161.
- Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 251, 467–470.
- Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645.
- Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raymackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. *Biochem. Biophys. Res. Commun.* 218, 777–782.
- Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. *Biochem. Pharmacol.* 51, 253–258.
- Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. *Curr. Biol.* 6, 1543–1544.

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,¹ Michael Bittner,² Jeffrey Trent,² J. Carl Barrett,¹ and Cynthia A. Afshari¹

¹Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

²Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog.* 24:153-159, 1999. © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae*, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cerevisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNase protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10-12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluorors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease-related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cerevisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22-24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25-27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28-30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only $4n$ cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)⁺ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and *BRCA1* [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on rela-

tive fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

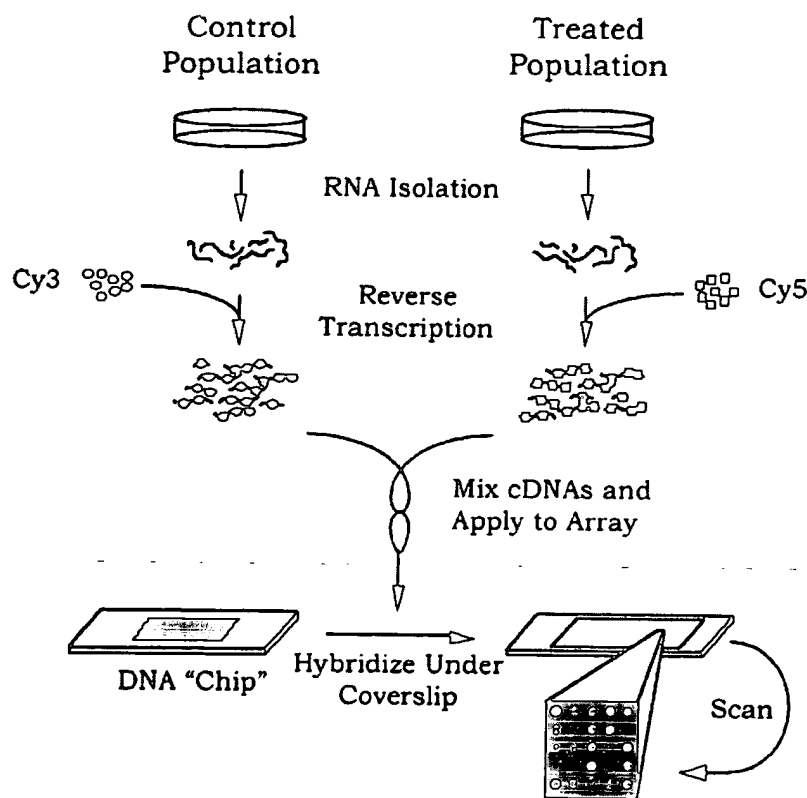


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illus-

trative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

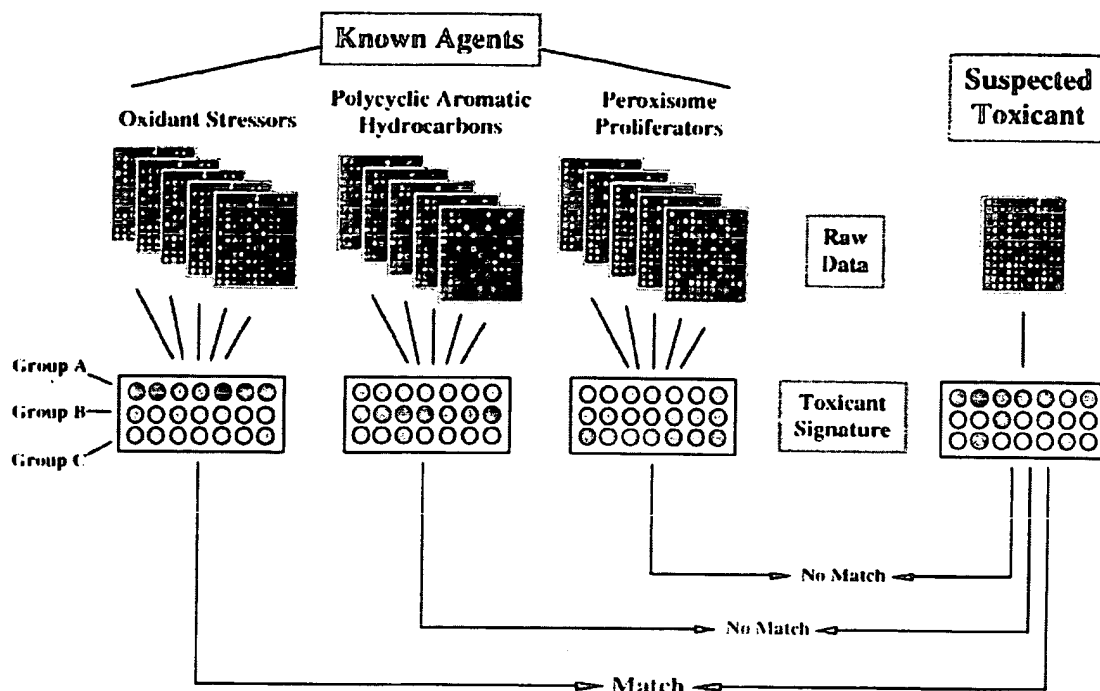


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce

consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures

are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

* This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive *in vivo* test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia*, and *Arabidopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under

which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

REFERENCES

1. <http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>
2. <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496-512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science* 1996;274:546, 563-567.
5. <http://www.perkin-elmer.com/press/prc5448.html>
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992;257:967-971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996;6:492-503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995;156:207-213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484-487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. *Science* 1995;270:467-470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14:457-460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* 1997;15:1359-1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. *Nat Biotechnol* 1998;16:27-31.
14. <http://www.synteni.com>
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;6:639-645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* 1997;2:364-374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58:5009-5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996; 93:10614-10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997;94:13057-13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 1997;94:2150-2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680-686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* 1996;37:29-40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the *E. coli* genome. *Genomics* 1996;37:77-86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *Biotechniques* 1994;17:328-329, 332-336.
25. <http://www.resgen.com/>
26. <http://www.genomesystems.com/>
27. <http://www.clontech.com/>
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 1994;91:5022-5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci USA* 1996;93:13555-13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 1995;19:442-447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675-1680.
34. <http://www.mdyn.com/>
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. *Genomics* 1996;33:445-456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. *Science* 1996;274:610-614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nat Genet* 1998;18:155-158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum Mutat* 1996;7:244-255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* 1996;14:441-447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* 1996;2:753-759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077-1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. *Science* 1998;281:1194-1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. *Environ Health Perspect* 1990;86:313-321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. *Nat Genet* 1998;20:19-23.
45. <http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html>
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;382:722-725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. *J Natl Cancer Inst* 1993;85:1159-1164.
48. <http://www.niehs.nih.gov/envgenom/home.html>

Subject: RE: [Fwd: Toxicol gy Chip]**Date: Mon. 3 Jul 2000 08:09:45 -0400****From: "Afshari.Cynthia" <afshari@niehs.nih.gov>****To: "Diana Hamlet-Cox" <dianahc@incyte.com>**

You can see the list of clones that we have on our 12K chip at:

<http://marvel.niehs.nih.gov/maps/guest/clonesrch.cfm>

We selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80-) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80- genes across our experiments.

Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be.

I hope this answers your question.

Cindy Afshari

> -----

> From: Diana Hamlet-Cox
> Sent: Monday, June 26, 2000 8:52 PM
> To: afshari@niehs.nih.gov
> Subject: [Fwd: Toxicology Chip]

> Dear Dr. Afshari,

> Since I have not yet had a response from Bill Grigg, perhaps he was not the right person to contact.

> Can you help me in this matter? I don't need to know the sequences, necessarily, but I would like very much to know what types of sequences are being used, e.g., GPCRs (more specific?), ion channels, etc.

> Diana Hamlet-Cox

> ----- Original Message -----

> Subject: Toxicology Chip
> Date: Mon. 19 Jun 2000 18:31:48 -0700
> From: Diana Hamlet-Cox <dianahc@incyte.com>
> Organization: Incyte Pharmaceuticals
> To: grigg@niehs.nih.gov

> Dear Colleague:

> I am doing literature research on the use of expressed genes as pharmacotoxicology markers, and found the Press Release dated February 29, 2000 regarding the work of the NIEHS in this area. I would like to know if there is a resource I can access (or you could provide?) that would give me a list of the 12,000 genes that are on your Human ToxChip Microarray. In particular, I am interested in the criteria used to select sequences for the ToxChip, including any control sequences included in the microarray.

> Thank you for your assistance in this request.

> Diana Hamlet-Cox, Ph.D.
> Incyte Genomics, Inc.

> --

> =====

> This email message is for the sole use of the intended recipient(s) and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited. If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.

> *****

>

>